USE OF AUXILIARY INFORMATION IN SAMPLING

DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE AWARD OF THE DEGREE OF

Master of Philosophy

IN

STATISTICS

By

TRAPTI THAKUR

Under the supervision of

Prof. M.Z. Khan

DEPARTMENT OF STATISTICS & OPERATIONS RESEARCH
ALIGARH MUSLIM UNIVERSITY
ALIGARH (INDIA)

2008

DEPARTMENT OF
STATISTICS & OPERATIONS RESEARCH
ALIGARH MUSLIM UNIVERSITY,
ALIGARH - 202002, INDIA.

Certificate

This is to certify that ***Mrs. Trapti Thakur*** *has carried out the work reported in the present dissertation entitled* ***"Use of Auxiliary Information in Sampling"*** *under my supervision. The dissertation is her own work and I recommend it for consideration for the award of* ***Master of Philosophy*** *in* ***Statistics.***

Dated: 2·8·08

Prof. M.Z. Khan
(Supervisor)

Dedicated to

My

Parents

And

Loving Husband

ACKNOWLEDGEMENT

I want to thank to 'GOD', the Almighty, the most powerful and the merciful whose blessings enabled me to complete this work in the present form.

I feel great pleasure at the completion of this work and would like to show my sincere gratitude to all the special persons who helped me.

It is, of course, my privilege and pleasure to express my profound sense of gratitude to my supervisor ***Prof. M. Z. Khan,*** *Department of Statistics and Operations Research, A.M.U., Aligarh for not only being dedicated and preserving but also his guidance, which seldom do justice to the latent qualities inherent in him. I must appreciate him for his patience, great involvement and sympathetic behavior, which enabled me to complete this work within stipulated time.*

I would like to thank ***Prof. A. H. Khan, Chairman, Department of Statistics and Operations Research A.M.U.****, Aligarh for his concern.*

I would like to thank to my teachers Prof. M. J. Ahsan, Prof. M. M. Khalid, Dr. A. Bari, Dr. M. Yaqub, Dr. R. U. Khan and Dr. Haseeb Athar.

I would like to thank to all non-teaching staff members of the department, for their kind help and cooperation.

I am greatly indebted to many of my research colleagues and friends who have encouraged me throughout this work. I am particularly grateful to Ms. Yojna, Ms. Shashi Saxena, Ms. Shazia Zarrin, Mr. Devendra Kumar, Mr. Rahul Varshney and Ms. Neeru for their constant encouragement and sharing of ideas during the preparation of manuscript.

My heart goes out in reverence to my parents, husband, brother and family members for their tremendous patience, forbearance, endurance and affection. All my appreciation for their support will not be adequate and enough to match their good wishes.

Trapti Thakur

Date: 2.8.08

Trapti Thakur

Department of Stats and O.R.

CONTENTS

PREFACE

This dissertation entitled "Use of Auxiliary Information in Sampling" is submitted to the Aligarh Muslim University, Aligarh, India, for the partial fulfilment of the degree of Master of Philosophy in Statistics. It embodies literature survey work carried out by me in the Department of Statistics and Operations Research, Aligarh Muslim University, Aligarh, India.

This dissertation consists of five chapters.

Chapter 1 deals with the basic ideas of sampling theory. It describes some basic concepts, and concerning results, which are relevant to the later chapters.

Chapter 2 deals with the problem of optimum stratification. For stratified sampling to be efficient the strata should be as homogenous as possible with respect to the study variable. In this chapter we consider the problem of optimum stratification when the information on the auxiliary variable x is used to estimate the populations mean Y of study variable y using ratio and regression method of estimation.

Chapter 3 deals with the problem of ratio estimate in successive sampling. In this chapter, successive sampling design is examined where a ratio estimate is used on the matched portion of the sample. A

comparison is made between this estimate and one which uses a regression estimate on the matched portion.

Chapter 4 deals with the problem of double sampling for ratio and regression estimator. In this chapter, proposed some ratio and regression estimators of the population mean using the Hansen and Hurwitz (1946) procedure is discussed. This procedure employs sub-sampling of the non-respondents assuming that the population mean of the auxiliary character is known.

Chapter 5 deals with the problem of post-stratification. Post-stratification could refer to any method of data analysis which involves forming units into homogenous groups after observation of the sample. A particularly useful paper due to Shukla and Dubey on PSNR (Post Stratified Non Response) has been discussed in some detail. This paper takes into account the use of auxiliary information and suggests a new estimator under the PSNR sampling scheme.

We have not discussed the programming approach in the body of the dissertation because we are not comfortable in the area of mathematical programming. We just wish to mention here that this is also a powerful tool for handling stratification problems with or without auxiliary variable.

A comprehensive list of references, arranged in alphabetical order is also provided at the end of dissertation.

CHAPTER I

PRELIMINERIES AND BASIC RESULTS

1.1 Sample and Sampling

Sample: Sample is a part or fraction of a population selected on some basis. Sample consists of a few items of a population. In principal a sample should be such that it is a true representative of the population.

Sampling: Sampling is the manner or scheme through which the required number of units is selected in a sample from a population.

1.2 Random or Probability Sampling

Probability sampling is the scientific method of selecting samples according to some laws of chance in which each unit in the population has some definite pre-assigned probability of being selected in the sample.

A sampling procedure which satisfies the following properties is termed as Random or Probability sampling:

i) There are number of samples of specified types $S_1, S_2, ..., S_k$ that can be formed by grouping units of a given population.

ii) Each possible sample S_i is assigned a known probability of selection p_i.

iii) The sampling procedure is capable of selecting any one of the possible sample S_i with probability p_i.

iv) The estimate constructed from any specific sample must be unique.

1.3 Simple Random Sampling

This is the method of selecting a sample of size n out of a finite population of size N in which each of the possible distinct samples has an equal chance of being selected is called simple random sampling.

We may have two distinct types of simple random sampling as follows:

i) Simple Random Sampling with Replacement (SRSWR)

ii) Simple Random Sampling without Replacement (SRSWOR)

In the following we give symbols, which are commonly used:

Y_i, value of the i^{th} unit of the population

y_i, value of the i^{th} unit of the sample

$$Y = \sum_{i=1}^{N} Y_i ,$$ population total

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i ,$$ population mean

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i ,$$ sample mean

$$f = \frac{n}{N} ,$$ sampling fraction

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N} (Y_i - \bar{Y})^2 ,$$ population mean square

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2 ,$$ sample mean square

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2, \qquad \text{population variance}$$

Symbols used in Sampling for Proportion

$$Y = \sum_{i=1}^{N} Y_i = NP = A, \qquad \text{population total}$$

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{A}{N} = P, \qquad \text{population mean}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - P)^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - P^2 = \frac{NP}{N} - P^2$$

$$= P(1-P) = PQ, \qquad \text{population variance}$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - P)^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - NP^2\right)$$

$$= \frac{1}{N-1}(NP - NP^2) = \frac{NP}{N-1}(1-P) = \frac{NPQ}{N-1}, \text{ population mean square}$$

Similarly,

$$y = \sum_{i=1}^{n} y_i = np = a, \qquad \text{sample total}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{a}{n} = p, \qquad \text{sample mean}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - p)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - np^2\right)$$

$$= \frac{1}{n-1}(np - np^2) = \frac{np}{n-1}(1-p) = \frac{npq}{n-1}, \quad \text{sample mean square}$$

Now we give some well known results without proof:

Theorem 1.3.1 In SRSWR, the sample mean $\bar{y}$ is an unbiased estimate of the population mean $\bar{Y}$ i.e. $\mathrm{E}(\bar{y}) = \bar{Y}$, and its variance is

$$V(\bar{y}) = \frac{N-1}{nN} S^2 = \frac{\sigma^2}{n}. \tag{1.3.1}$$

Theorem 1.3.2 In SRSWR, the sample mean square s^2 is an unbiased estimate of the population variance σ^2 i.e. $E(s^2) = \sigma^2 = \left(\frac{N-1}{N}\right) S^2$.

Corollary 1.3.1 In SRSWR, the estimated population total $\hat{Y} = N\bar{y}$ is an unbiased estimate of the population total Y i.e. $E(\hat{Y}) = Y$, and its variance is

$$V(\hat{Y}) = \frac{N^2\sigma^2}{n}. \tag{1.3.2}$$

Theorem 1.3.3 In SRSWR, the sample proportion $p = \frac{a}{n}$ is an unbiased estimate of the population proportion $P = \frac{A}{N}$ i.e. $E(p) = P$, and the variance of p is

$$V(p) = \frac{PQ}{n}. \tag{1.3.3}$$

Theorem 1.3.4 In SRSWOR, the sample mean $\bar{y}$ is an unbiased estimate of the population mean $\bar{Y}$ i.e. $E(\bar{y}) = \bar{Y}$, and its variance is

$$V(\bar{y}) = \left(\frac{N-n}{nN}\right)S^2 = \left(\frac{1}{n} - \frac{1}{N}\right)S^2 = (1-f)\frac{S^2}{n}. \qquad (1.3.4)$$

Theorem 1.3.5 In SRSWOR, the sample mean square s^2 is an unbiased estimate of the population mean square S^2 i.e. $E(s^2) = S^2$.

Corollary 1.3.2 In SRSWOR, the estimated population total $\hat{Y} = N\bar{y}$ is an unbiased estimate of the population total Y i.e. $E(\hat{Y}) = Y$, and its variance is

$$V(\hat{Y}) = N^2(1-f)\frac{S^2}{n}. \qquad (1.3.5)$$

Theorem 1.3.6 In SRSWOR, the sample proportion $p = \frac{a}{n}$ is an unbiased estimate of the population proportion $P = \frac{A}{N}$ i.e. $E(p) = P$, and the variance of p is

$$V(p) = \left(\frac{N-n}{N-1}\right)\frac{PQ}{n}. \qquad (1.3.6)$$

Property: $V(\bar{y})$ under SRSWOR is less than the $V(\bar{y})$ under SRSWR,

i.e. $V(\bar{y})\text{SRSWOR} < V(\bar{y})\text{SRSWR}$

$$V(\bar{y})_{SRSWOR} = \left(\frac{N-n}{nN}\right)S^2 < V(\bar{y})_{SRSWR} = \left(\frac{N-1}{nN}\right)S^2$$

Hence, SRSWOR provides more efficient estimator of $\bar{Y}$ relative to SRSWR.

1.4 Stratified Random Sampling

Apart from increasing the sample size, one possible way to estimate the population mean or total with greater precision is to divide the population

in several groups (sub-population or classes, these sub-populations are non-overlapping and are called strata) each of which is more homogenous than the entire population and then draw a random sample of predetermined size from each one of the groups. The groups, into which the population is divided, are called strata or each group is called stratum. The whole procedure of dividing the population into the strata and then drawing a random sample from each of the strata is called stratified random sampling.

The use of stratified sampling in sample survey needs the solution of the following three basic problems:

i) The determination of the number of strata

ii) The determination of the strata boundaries

iii) The determination of the sizes of the samples to be selected from various strata.

Let the population of size N be divided into k strata of sizes $N_1, N_2, ..., N_k$. These strata are mutually exclusive (non-overlapping) such that $N_1 + N_2 + ... + N_k = \sum_{i=1}^{k} N_i = N.$

For full benefit from stratification the sub-population sizes, $N_i (i = 1,2,...,k)$, must be known. Furthermore, let a sample of sizes $n_1, n_2, ..., n_k$, be drawn (by the method of SRS) from each group (stratum) independently, the sample size within i^{th} stratum being n_i, such that

$$n_1 + n_2 + ... + n_k = \sum_{i=1}^{k} n_i = n.$$

The suffix i denotes the stratum and j the unit within stratum. The following symbols all refer to stratum i.

N_i , total number of units

n_i , number of units in sample

$f_i = \frac{n_i}{N_i}$, sampling fraction in the stratum

$w_i = \frac{n_i}{n}, W_i = \frac{N_i}{N}$, stratum weight

y_{ij}, value of the characteristic under study for the i^{th} unit

$Y_i = \sum_{j=1}^{N_i} Y_{ij}$, total based on N_i units (stratum total)

$\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$, mean based on N_i units (stratum mean)

$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$, mean based on n_i units (sample mean)

$\sigma_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (y_{ij} - \bar{Y}_i)^2$, variance based on N_i units (stratum variance)

$S_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (y_{ij} - \bar{Y}_i)^2$,mean square based on N_i units (stratum mean square)

$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$, sample mean square based on n_i units

$$Y = \sum_{i=1}^{k}\sum_{j=1}^{N_i} y_{ij} = \sum_{i=1}^{k} N_i \bar{Y}_i, \quad \text{population total}$$

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} y_{ij} = \sum_{i=1}^{k} W_i \bar{Y}_i, \quad \text{overall population mean}$$

$$\hat{Y}_i = \frac{N_i}{n_i}\sum_{j=1}^{n_i} y_{ij}, \qquad \text{estimated total}$$

The estimate of the population mean per unit, used in stratified sampling is denoted by $\bar{y}_{st}$ and is given by

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i = \sum_{i=1}^{k} W_i \bar{Y}_i, \tag{1.4.1}$$

$\bar{y}_{st}$ is not, in general, the same as the mean $\bar{y}$,

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{y}_i$.

both coincides if in every stratum $\frac{n_i}{n} = \frac{N_i}{N} = \frac{n_i}{N_i} = \frac{n}{N}$ or $f_i = f$,

that is if sampling fraction is the same for all strata. This stratification is known as stratification with proportional allocation of n_i.

Theorem 1.4.1 For stratified random sampling, without replacement, if in every stratum the sample estimate $\bar{y}_i$ is an unbiased estimate of

$\bar{Y}_i$,(i.e. $E(\bar{y}_i) = \bar{Y}_i$), and samples are drawn independently in different strata, then $\bar{y}_{st}$ is an unbiased estimate of the overall population mean $\bar{Y}$, that is $E(\bar{y}_{st}) = \bar{Y}$, and its variance is

$$V(\bar{y}_{st}) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) W_i^2 S_i^2 = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} (1 - f_i) \qquad (1.4.2)$$

We see that variance of $\bar{y}_{st}$ depends on S_i^2, the heterogeneity within the strata. Thus, if S_i^2 are small (strata are homogeneous) then the stratified random sampling provides estimates with greater precision.

Corollary 1.4.1 For stratified random sampling, without replacement, the estimated population total $\hat{Y}_{st} = N \bar{y}_{st}$ is an unbiased estimate of the population total Y i.e. $E(\hat{Y}_{st}) = Y$, and its variance is

$$V(\hat{y}_{st}) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) N_i^2 S_i^2 = \sum_{i=1}^{k} \frac{N_i^2 S_i^2}{n_i} (1 - f_i) \qquad (1.4.3)$$

Remarks:

(a) If in every stratum $\frac{n_i}{n} = \frac{N_i}{N}$, the variance of $\bar{y}_{st}$ reduces to

$$V(\bar{y}_{st}) = \frac{1}{n} \left(\frac{N-n}{N} \right) \sum_{i=1}^{k} W_i S_i^2 = \frac{1-f}{n} \sum_{i=1}^{k} W_i S_i^2 \ .$$

(b) If in every stratum $\frac{n_i}{n} = \frac{N_i}{N}$, and the variance of $\bar{y}_{st}$ in all strata have the same value S^2, then the result reduces to

$$V(\bar{y}_{st}) = \frac{1}{n}\left(\frac{N-n}{N}\right)S^2 \sum_{i=1}^{k} W_i = \frac{1-f}{n}S^2, \quad \text{since} \quad \sum_{i=1}^{k} W_i = 1$$

(c) If N_i are large as compared to n_i (that is if the sampling fractions f_i are negligible in all strata), then

(i) $V(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i}$

(ii) $V(\hat{y}_{st}) = \sum_{i=1}^{k} \frac{N_i^2 S_i^2}{n_i}$.

Theorem 1.4.2 If stratified random sampling is with replacement, then $\bar{y}_{st}$ is an unbiased estimate of population mean $\bar{Y}$, that is $E(\bar{y}_{st}) = \bar{Y}$, and its variance is

$$V(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} \tag{1.4.4}$$

Corollary 1.4.2 For stratified random sampling, with replacement, the estimated population total $\hat{Y}_{st} = N\,\bar{y}_{st}$ is an unbiased estimate of the population total Y i.e. $E(\hat{Y}_{st}) = Y$, and its variance is

$$V(\hat{Y}_{st}) = N^2 V(\bar{y}_{st}) = N^2 \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} = \sum_{i=1}^{k} \frac{N_i^2 S_i^2}{n_i} \tag{1.4.5}$$

Theorem 1.4.3 If a simple random sample is taken within each stratum, then an unbiased estimate of S_i^2 is s_i^2, and an unbiased estimate of variance $\bar{y}_{st}$ is given by

$$\hat{V}(\bar{y}_{st}) = v(\bar{y}_{st}) = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)W_i^2 s_i^2 = \sum_{i=1}^{k} \frac{W_i^2 s_i^2}{n_i}(1-f_i) \tag{1.4.6}$$

1.5 Principal Reasons for Stratification

The following points regarding stratification to be noted:

1) To gain in precision, divide a heterogeneous population into strata in such a way that each stratum is internally homogeneous.

2) To accommodate administrative convenience and or cost considerations, fieldwork is organized by strata, which usually results in saving in cost and effort.

3) To obtain separate estimates for strata.

4) We can accommodate different sampling plan in different strata.

5) We can have data of known precision for certain subdivisions treating each subdivision as a "population" in its own right.

6) There may be marked difference in sampling problems in different parts of the population.

1.6 Allocation of Sample to Different Strata

An important consideration is how to allocate a total sample size n among the k identified strata. There are three methods of allocation of sample sizes to different strata in a stratified random sampling.

1.6.1 Equal Allocation

If the strata are presumed to be of roughly equal size, and there is no additional information regarding the variability or distribution of the response in the strata, equal allocation to the strata is probably the best choice:

$$n_i = \frac{n}{k} \tag{1.6.1}$$

and its variance is

$$V(\bar{y}_{st})_{equal} = \frac{k}{n}\sum_{i=1}^{k} W_i^2 S_i^2 - \frac{1}{N}\sum_{i=1}^{k} W_i S_i^2 \tag{1.6.2}$$

1.6.2 Proportional Allocation

This allocation generally known as proportional allocation was originally proposed by Bowley. When no other information except N_i, the total number of units in the i^{th} stratum, is available, the allocation of a given sample of size n to different strata is done in proportion to their sizes, i.e. in the i^{th} stratum $n_i \propto N_i$ or $n_i = \lambda N_i$, where λ is the constant of proportionality, and

$$\sum_{i=1}^{k} n_i = \lambda \sum_{i=1}^{k} N_i \text{ ,or } \lambda = \frac{n}{N}, \quad \Rightarrow n_i = \frac{n}{N} N_i = nW_i \tag{1.6.3}$$

and its variance is

$$V(\bar{y}_{st})_{prop} = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i S_i^2 = \frac{1-f}{n}\sum_{i=1}^{k} W_i S_i^2 \tag{1.6.4}$$

1.6.3 Optimum Allocation

The formula for optimum allocation in various strata was derived by Tschuprow in (1923). Later J. Neyman derived them independently in (1934). That is why such an allocation is often termed Neyman-Tschuprow allocation. In this method of allocation the sample sizes n_i in the respective strata are determined with a view to minimize $V(\bar{y}_{st})$ for a

specified cost of conducting the sample survey or to minimize the cost for a specified value of $V(\bar{y}_{st})$.

The simplest cost function is of the form

$$\text{Cost} = C = c_0 + \sum_{i=1}^{k} c_i n_i$$

where the overhead cost c_0 is constant and c_i is the average cost of surveying one unit in the i^{th} stratum, then

$$C - c_0 = \sum_{i=1}^{k} c_i n_i = C'(say) \tag{1.6.5}$$

and, we know that

$$V(\bar{y}_{st}) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) W_i^2 S_i^2 = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} - \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{N_i}$$

$$\text{Thus } V(\bar{y}_{st}) + \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{N_i} = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} = V'(say) \tag{1.6.6}$$

where C' and V' are the functions of n_i.

To determine n_i such that $V(\bar{y}_{st})$ is minimum and cost

$C = c_0 + \sum_{i=1}^{k} c_i n_i$ is fixed, consider the function

$$\phi = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} - \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{N_i} + \lambda (c_0 + \sum_{i=1}^{k} c_i n_i - C)$$

where λ is some unknown constant.

Using the calculus method of Lagrange multipliers, we select n_i, and the constant λ to minimize ϕ. Differentiating ϕ with respect to n_i, and equating to zero, we have

$$\frac{\partial \phi}{\partial n_i} = 0 = -\frac{W_i^2 S_i^2}{n_i^2} + \lambda c_i \text{ or } n_i = \frac{1}{\sqrt{\lambda}} \frac{W_i S_i}{\sqrt{c_i}} \tag{1.6.7}$$

$$\Rightarrow n_i \propto \frac{W_i S_i}{\sqrt{c_i}} \quad or \quad n_i \propto \frac{N_i S_i}{\sqrt{c_i}}$$

This allocation is known as optimum allocation.

Taking summation on both the sides of equation (1.7.7), we get-

$$\sum_{i=1}^{k} n_i = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{k} \frac{W_i S_i}{\sqrt{c_i}} \quad \text{or} \quad \frac{1}{\sqrt{\lambda}} = \frac{n}{\sum_{i=1}^{k} \frac{W_i S_i}{\sqrt{c_i}}}$$

$$\Rightarrow n_i = \frac{n}{\sum_{i=1}^{k} \frac{W_i S_i}{\sqrt{c_i}}} \frac{W_i S_i}{\sqrt{c_i}} = n \frac{W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}} = n \frac{N_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} N_i S_i / \sqrt{c_i}} \tag{1.6.8}$$

The total sample size n required for optimum the sample sizes within strata. The solution for the value of n depends on whether the sample is chosen to meet a specified total cost C or to give a specified variance V for $\bar{y}_{st}$.

(i) If cost is fixed: Substitute the optimum values of n_i in cost equation (1.6.5) and solve for n as,

$$C - c_0 = \sum_{i=1}^{k} c_i n_i = \sum_{i=1}^{k} n \frac{W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}} c_i = n \sum_{i=1}^{k} \frac{W_i S_i \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}}$$

$$\Rightarrow n = \frac{(C - c_0) \sum_{i=1}^{k} W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i \sqrt{c_i}} \tag{1.6.9}$$

The value of n_i is obtained after substituting this value of n in equation (1.6.8), we get

$$n_i = \frac{(C - c_0) \sum_{i=1}^{k} W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i \sqrt{c_i}} \times \frac{W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i / \sqrt{c_i}}$$

$$n_i = \frac{(C - c_0) W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} W_i S_i \sqrt{c_i}} \tag{1.6.10}$$

$V(\bar{y}_{st})$ under optimum allocation for fixed cost:

$$V(\bar{y}_{st})_{opt} = \sum_{i=1}^{k} \left[\frac{\sum_{i=1}^{k} W_i S_i \sqrt{c_i}}{(C - c_0) W_i S_i / \sqrt{c_i}} - \frac{1}{N_i} \right] W_i^2 S_i^2$$

$$= \sum_{i=1}^{k} \left[\frac{\left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \right) W_i^2 S_i^2}{(C - c_0) W_i S_i / \sqrt{c_i}} - \frac{W_i^2 S_i^2}{N_i} \right]$$

$$= \sum_{i=1}^{k} \left[\frac{1}{(C - c_0)} \left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \right) W_i S_i \sqrt{c_i} - \frac{1}{N_i} \left(\frac{N_i}{N} \right) W_i S_i^2 \right]$$

$$V(\bar{y}_{st})_{opt} = \left[\frac{1}{(C - c_0)} \left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \right)^2 - \frac{1}{N} \sum_{i=1}^{k} W_i S_i^2 \right] \qquad (1.6.11)$$

2) If the variance is fixed: Substituting the optimum n_i from equation (1.6.8) in equation (1.6.6), we get

$$V(\bar{y}_{st}) + \frac{1}{N} \sum_{i=1}^{k} W_i S_i^2 = \sum_{i=1}^{k} \frac{W_i^2 S_i^2}{n_i} = \sum_{i=1}^{k} \frac{W_i^2 S_i^2 \sum_{i=1}^{k} W_i S_i / \sqrt{c_i}}{n W_i S_i / \sqrt{c_i}}$$

$$= \frac{1}{n} \sum_{i=1}^{k} W_i S_i \sqrt{c_i} \left(\sum_{i=1}^{k} W_i S_i / \sqrt{c_i} \right)$$

$$\text{Thus} \quad n = \frac{\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \left(\sum_{i=1}^{k} W_i S_i / \sqrt{c_i} \right)}{V + \frac{1}{N} \sum_{i=1}^{k} W_i S_i^2} \qquad (1.6.12)$$

and then the value of n_i is obtained by putting this value of n in equation (1.6.8)

$$n_i = \frac{\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \left(\sum_{i=1}^{k} W_i S_i / \sqrt{c_i} \right)}{V + \frac{1}{N} \sum_{i=1}^{k} W_i S_i^2} \times \frac{W_i S_i / \sqrt{c_i}}{\sum_{i=1}^{k} (W_i S_i / \sqrt{c_i})}$$

$$\Rightarrow n_i = \frac{(W_i S_i / \sqrt{c_i}) \sum_{i=1}^{k} W_i S_i \sqrt{c_i}}{V + \frac{1}{N} \sum_{i=1}^{k} W_i S_i^2} \qquad (1.6.13)$$

Optimum Cost for fixed Variance: We know that $C - c_0 = \sum_{i=1}^{k} c_i n_i$,

substituting the value of n_i from equation (1.7.13), we get

$$C - c_0 = \sum_{i=1}^{k} c_i \left(\frac{(W_i S_i / \sqrt{c_i}) \sum_{i=1}^{k} W_i S_i \sqrt{c_i}}{V + \frac{1}{N} \sum_{i=1}^{k} W_i S_i^2} \right)$$

$$= \sum_{i=1}^{k} \frac{W_i S_i \sqrt{c_i} \left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \right)}{V + \frac{1}{N} \sum_{i=1}^{k} W_i S_i^2} = \frac{\left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \right)^2}{V + \frac{1}{N} \sum_{i=1}^{k} W_i S_i^2}$$

$$\Rightarrow C = c_0 + \frac{\left(\sum_{i=1}^{k} W_i S_i \sqrt{c_i} \right)^2}{V + \frac{1}{N} \sum_{i=1}^{k} W_i S_i^2} \qquad (1.6.14)$$

Remark:

An important special case arises if $c_i = c$, that is, if the cost per unit is the same in all the strata. The cost becomes $C = c_0 + \sum_{i=1}^{k} c_i n_i = c_0 + cn$, and optimum allocation for fixed cost reduces to optimum allocation for fixed sample size. In this case $V(\bar{y}_{st})$ is minimized for a fixed total size of sample n if

$$n_i = n \frac{W_i S_i}{\sum_{i=1}^{k} W_i S_i} = n \frac{N_i S_i}{\sum_{i=1}^{k} N_i S_i} \qquad (1.6.15)$$

$$\Rightarrow n_i \propto W_i S_i \quad \text{or} \quad n_i \propto N_i S_i \ ,$$

this allocation is called the Neyman or Neyman-Schupure allocation.

$V(\bar{y}_{st})$ under optimum allocation for fixed n:

$$V(\bar{y}_{st})_{opt} = \sum_{i=1}^{k}\left(\frac{1}{nW_iS_i}\sum_{i=1}^{k}W_iS_i - \frac{1}{N_i}\right)W_i^2S_i^2$$

$$= \sum_{i=1}^{k}\left[\frac{1}{nW_iS_i}\left(\sum_{i=1}^{k}W_iS_i\right)W_i^2S_i^2 - \left(\frac{1}{N_i}\right)W_i^2S_i^2\right]$$

$$= \sum_{i=1}^{k}\left[\frac{1}{n}\left(\sum_{i=1}^{k}W_iS_i\right)W_iS_i - \frac{1}{N_i}\left(\frac{N_i}{N}\right)W_iS_i^2\right]$$

$$V(\bar{y}_{st})_{opt} = \frac{1}{n}\left(\sum_{i=1}^{k}W_iS_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k}W_iS_i^2 \qquad (1.6.16)$$

Note: If N is large, $V(\bar{y}_{st})_{opt}$ reduces to $V(\bar{y}_{st})_{opt} = \frac{1}{n}\left(\sum_{i=1}^{k}W_iS_i\right)^2$.

1.7 Ratio Estimate

In the ratio method an auxiliary variate x_i, correlated with y_i, is obtained for each unit in the sample. The population total X of the x_i must be known. In practice, x_i is often the value of y_i at same previous time when a complete census was taken. The aim in this method is to obtain increased precision by taking advantage of the correlation between y_i and x_i. At present we assume simple random sampling.

The ratio estimate of Y, the population total of the y_i, is

$$\hat{Y}_R = \frac{y}{x}X = \frac{\bar{y}}{\bar{x}}X \qquad (1.7.1)$$

The ratio estimate of $\bar{Y}$, the population mean of the y_i is

$$\hat{\bar{Y}}_R = \frac{y}{x}\bar{X} \qquad (1.7.2)$$

Notations

y_i: Measurement of the main variable on the i^{th} unit of the population.

x_i : Measurement of the auxiliary variable on the i^{th} unit of the population.

$Y = \sum_{i=1}^{N} y_i$, Population total of y

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{Y}{N}$, Population mean of y

$X = \sum_{i=1}^{N} x_i$, Population total of x

$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{X}{N}$, Population mean of x

$R = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}}$, Ratio of the population totals or means of y and x

ρ , Correlation coefficients between y and x in the population

Theorem1.7.1 In SRSWOR, for large $n, \hat{R} = \frac{\bar{y}}{\bar{x}}$ is approximately unbiased for the population ratio R and has an approximate variance

$$V(\hat{R}) = \frac{1-f}{n(N-1)\bar{X}^2} \sum_{i=1}^{N} (y_i - R x_i)^2 \qquad (1.7.3)$$

Alternative Expressions Of $V(\hat{R})$:

(1) **In Terms f Correlation Coefficient:**

The correlation coefficient ρ between y and x is defined by

$$\rho = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})(x_i - \bar{X})}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{Y})^2 \sum_{i=1}^{N} (x_i - \bar{X})^2}} = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})(x_i - \bar{X})}{(N-1) S_y S_x},$$

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{X})^2$$

OR

$$\sum_{i=1}^{N} (y_i - \bar{Y})(x_i - \bar{X}) = (N-1)\rho S_y S_x$$

$$V(\hat{R}) = \frac{1-f}{n} R^2 [C_{yy} + C_{xx} - 2\rho C_y C_x]$$

Where $C_y = \frac{S_y}{\bar{Y}}$ and $C_x = \frac{S_x}{\bar{X}}$ are the coefficient of variation of y and x respectively. Thus C_{yy} and C_{xx} are the square of the coefficient of variation and are also called relative variances.

(2) **In Terms Of Covariance:**

The covariance of y and x is defined by

$$S_{yx} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X}) = \rho S_y S_x \text{. Thus,}$$

$$V(\hat{R}) = \frac{1-f}{n}R^2[C_{yy} + C_{xx} - 2C_{yx}]$$

where, C_{yx} is called relative covariance.

1.8 Comparison of Ratio Estimate with Mean per Unit Estimate

The conditions under which the ratio estimator is superior to the mean per unit will be worked out with a comparison of their variances.

In SRSWOR, the variance of the mean per unit is given by

$$V(\bar{y}) = (\frac{1}{n} - \frac{1}{N})S^2 = \frac{1-f}{n}S_y^2, \text{ and}$$

the variance of the mean based on the ratio method is given by

$$V(\hat{\bar{Y}}_R) = \frac{1-f}{n}[S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x]$$

Obviously ratio estimate $\hat{\bar{Y}}_R$ will more precise as compared to $\bar{y}$ if and only $V(\hat{\bar{Y}}_R) < V(\bar{y})$. So that

$$\frac{1-f}{n}[S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x] < \frac{1-f}{n}S_y^2$$

$$\Rightarrow \qquad R^2 S_x^2 < 2\rho R S_y S_x$$

$$\rho > \frac{R^2 S_x^2}{RS_y S_x} = \frac{RS_x}{2S_y} \quad \text{Or}$$

$$\rho > \frac{1}{2}\frac{S_{x/\bar{X}}}{S_y/\bar{Y}}, \text{ or} \qquad \rho > \frac{1}{2}\left(\frac{CV(x)}{CV(y)}\right)$$

Theorem1.7.2 In simple random sampling, the bias of the ratio estimator $\hat{R}$ is $B(\hat{R}) = -\frac{Cov(\hat{R}, \bar{x})}{\bar{X}}$

Theorem1.7.3 If y_i, x_i is the pair of variates defined on every unit in the population and $\bar{y}, \bar{x}$ are the corresponding means from a simple random sample of size n, then their covariance

$$E(\bar{y} - \bar{Y})(\bar{x} - \bar{X}) = \frac{N-n}{nN}\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X}).$$

Theorem1.7.4 Show that the first approximation to the relative bias of the ratio estimator in the simple random sampling, WOR, is given by

$$\frac{B(\hat{R})}{R} = \frac{1-f}{n\bar{Y}\bar{X}}(RS_x^2 - \rho S_y S_x) \cong \frac{1-f}{n}(C_{xx} - \rho C_x C_y)$$

where $C_x = \frac{S_x}{\bar{X}}$, and $C_y = \frac{S_y}{\bar{Y}}$ are the coefficients of variation of x and y, respectively.

1.9 Optimum Property of Ratio Estimator

With SRS from infinite population, the ratio estimate of the population mean $\bar{Y}$ is the best linear unbiased estimate if following two conditions are satisfied:

(1) The relation between x_i and y_i is straight line through origin.

(2) The variance of y_i about this line is proportional to x_i.

Proof:

(1) $E(y_i) = E(y_i/x_i) = Bx_i$, under this condition, we can write the model

$y_i = Bx_i + e_i$, where $E(e_i) = E(e_i/x_i) = 0$.

(2) $V(y_i) = V(Bx_i) + V(e_i/x_i) + Cov(Bx_i, e_i)$

$= V(e_i/x_i)$, covariance is zero, as Bx_i and e_i are

independent

$= Ax_i$

$\Rightarrow V(y_i) \propto x_i$, if and only if $V(e_i/x_i) = Ax_i$.

Let $b = \sum_{i=1}^{n} \alpha_i y_i$ is a linear estimate of B, then

$$E(b) = \sum_{i=1}^{n} \alpha_i E(y_i) = \sum_{i=1}^{n} \alpha_i E(y_i/x_i) = \sum_{i=1}^{n} \alpha_i Bx_i$$

$$= B\sum_{i=1}^{n} \alpha_i x_i = B, \text{ if and only if } \sum_{i=1}^{n} \alpha_i x_i = 1$$

Thus, the class $b = \sum_{i=1}^{n} \alpha_i y_i$ is the class of linear unbiased estimator of B,

if $\sum_{i=1}^{n} \alpha_i x_i = 1$.

$$V(b) = V\left(\sum_{i=1}^{n} \alpha_i y_i\right) = \sum_{i=1}^{n} \alpha_i^2 V(y_i) = \sum_{i=1}^{n} \alpha_i Ax_i = \lambda \sum_{i=1}^{n} \alpha_i^2 x_i.$$

Now our problem is to find the estimate in the class of estimators $b=\sum_{i=1}^{n}\alpha_i y_i$, such that $V(b)$ is minimum under the condition $\sum_{i=1}^{n}\alpha_i x_i=1$. For this we shall use the Lagrange multipliers technique.

Define the function

$$\phi=V(b)+L\left(1-\sum_i \alpha_i x_i\right)=A\sum_i \alpha_i^2 x_i+\lambda\left(1-\sum_i \alpha_i x_i\right)$$

To get α_i's such that $V(b)$ is minimum, subject to condition $\sum_{i=1}^{n}\alpha_i x_i=1$ is same as to get α_i's such that ϕ is minimum. Differentiate ϕ with respect to α_i and equate to zero, we get

$$\frac{\partial\phi}{\partial\alpha_i}=0=2A\alpha_i x_i-\lambda x_i \quad \text{or} \quad 2A\alpha_i x_i=\lambda x_i \quad \text{or} \quad \alpha_i=\frac{\lambda}{2A}$$

But

$$\sum_{i=1}^{n}\alpha_i x_i=1 \quad \text{or} \quad \sum_{i=1}^{n}\frac{\lambda x_x}{2A} \quad \text{or} \quad \frac{\lambda}{2A}=\frac{1}{\sum_i x_i} \quad \text{or} \quad \alpha_i=\frac{1}{\sum_i x_i}, \text{ for all } i.$$

Hence, the estimator b is linear class $b=\sum_{i=1}^{n}\alpha_i y_i$ has minimum $V(b)$ for $\alpha_i=\frac{1}{\sum_i x_i}$.

Thus, the best linear unbiased estimator of B is given by

$$b=\sum_{i=1}^{n}\alpha_i y_i=\frac{\sum_i y_i}{\sum_i x_i}=\hat{R}$$

Now arranging the modal over the population, we get $\overline{Y} = B\overline{X}$, as $b = \hat{R}$ is the best linear unbiased estimate of B. Therefore, the best linear unbiased estimate of $\overline{Y}$ will be $\hat{\overline{Y}} = b\overline{X} = \hat{R}\overline{X}$, which is the ratio estimate of $\overline{Y}$.

1.10 Ratio Estimators in Stratified Sampling

In stratified random sampling there are two ways of forming the ratio estimator of a population total. These give

(i) Separate ratio estimator

(ii) Combined ratio estimator

(i) Separate Ratio Estimator: If $\overline{y}_i$ and $\overline{x}_i$ are sample mean computed from the i^{th} stratum, the ratio $\frac{\overline{y}_i}{\overline{x}_i}$ are computed separately from each stratum and with the knowledge of population total X_i for the i^{th} stratum, we may define the estimator $\hat{Y}_{RS}$ (s for separate) as

$$\hat{Y}_{RS} = \sum_{i=1}^{k} \frac{\overline{y}_i}{\overline{x}_i} X_i \qquad (1.10.1)$$

Theorem1.10.1 If an independent simple random sample is drawn in each stratum and sample sizes are large in all strata, then

$$V(\hat{Y}_{RS}) = \sum_{i=1}^{k} \frac{N_i^2(1-f_i)}{n_i}(S_{yi}^2 + R_i^2 S_{xi}^2 - 2R_i\rho_i S_{yi}S_{xi}) \qquad (1.10.2)$$

where $R_i = \frac{\overline{Y}_i}{\overline{X}_i} = \frac{Y_i}{X_i}$ and ρ_i are the true ratio and the coefficient of correlation, respectively in the i^{th} stratum.

Corollary1.10.1 In stratified random sampling, WOR, almost unbiased estimator of $V(\hat{Y}_{Rs})$ is given by

$$\hat{V}(\hat{Y}_{Rs}) = \sum_{i=1}^{k} \frac{N_i^2(1-f_i)}{n_i}(s_{yi}^2 + \hat{R}_i^2 s_{xi}^2 - 2\hat{R}_i s_{yxi}) \qquad (1.10.3)$$

where s_{yxi} stands for the estimated covariance in the i^{th} stratum.

(ii) Combined Ratio Estimator: If X_i is not known, but only X is known, and if R_i's do not differ considerably from stratum to stratum and if $n_i's$ are large enough, one can use the combined ratio estimator $\hat{Y}_{Rc}$(c for combined) for a sample from a stratified population as

$$\hat{Y}_{Rc} = \frac{\hat{Y}_{st}}{\hat{X}_{st}} = \frac{\bar{y}_{st}}{\bar{x}_{st}} X, \qquad (1.10.4)$$

where $\bar{y}_{st} = \sum_{i=1}^{k} \frac{N_i}{N}\bar{y}_i = \frac{1}{N}\hat{Y}_{st}$, and $\bar{x}_{st} = \sum_{i=1}^{k} \frac{N_i}{N}\bar{x}_i = \frac{1}{N}\hat{X}_{st}$.

Theorem1.10.2 If the total sample size n is large and simple random sampling, WOR, is done in each stratum independently, then $\hat{Y}_{Rc}$ is a consistent estimator and its

$$V(\hat{Y}_{Rc}) = \sum_{i=1}^{k} \frac{N_i^2(1-f_i)}{n_i}(S_{yi}^2 + S_{xi}^2 - 2R\rho_i S_{yi} S_{xi}) \qquad (1.10.5)$$

Corollary1.10.2 In stratified random sampling, WOR, almost unbiased estimator of $V(\hat{Y}_{Rc})$ is given by

$$\hat{V}(\hat{Y}_{Rc}) = \sum_{i=1}^{k} \frac{N_i^2(1-f_i)}{n_i}(s_{yi}^2 + \hat{R}^2 s_{xi}^2 - 2\hat{R}s_{yxi}) \qquad (1.10.6)$$

Where s_{yxi} stands for the estimated covariance in the i^{th} stratum.

1.11 Regression Method of Estimation

Like ratio estimators, linear regression estimators also make use of auxiliary information (variable) that is correlated with under study variable for increasing precision. Ratio estimators often result in increased precision if the regression of under study variable (y) on auxiliary variable (x) is linear and passes through the origin i.e. when the regression equation of y on x is $y = bx$. When the regression of y on x is linear, the regression line does not pass through the origin. In such situations, it is better to use estimators based on linear regression.

Let a SRS of size n has been obtained from a population of size N and y_i and x_i measured on each unit of the sample and the population mean $\bar{X}$ of the x variate is known. The regression estimators of the population mean $\bar{Y}$ and population total Y are given by

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}) = \bar{y} - b(\bar{x} - \bar{X}), \text{ and } \hat{Y}_{lr} = N[\bar{y} + b(\bar{X} - \bar{x})].$$

Where, lr denotes linear regression, and b is an estimate of the regression coefficient B of y on x in the population (an estimate of the change in y for a unit change in x). The rationale (belief) of this estimate is that if $\bar{x}$ is below average we should expect $\bar{y}$ also to be below average by an amount $b(\bar{X} - \bar{x})$ because of the regression of y_i and x_i.

Watson, JD (1937) used a regression of leaf area on leaf weight to estimate the average area of the leaves on a plant. The procedure was to weight all the leaves on the plants. For a small sample of leaves, the area and the weight of each leaf were determined. The sample mean leaf area was than adjusted by means of regression on leaf weight. The point of the application is, of course, that the weight of a leaf can be found quickly but determination of its area is more time consuming. In another application described by Yates, Z (1960), an eye estimate of the volume of timber was made on each of a population of 1/10 acre plots, and the actual timber volume was measured for a sample of plots. The regression estimates adjust the sample mean of the actual measurements on the rapid estimates.

Theorem1.11.1 In simple random sampling, WOR, in which b_0 is a pre-assigned constants, the linear regression estimate $\bar{y}_{lr} = \bar{y} + b_0(\bar{X} - \bar{x})$ is an unbiased estimate of $\bar{Y}$ with its variance

$$V(\bar{y}_{lr}) = \frac{1-f}{n(N-1)} \sum_{i=1}^{N} [(y_i - \bar{Y}) - b_0(x_i - \bar{X})]^2 = \frac{1-f}{n} [S_y^2 + {b_o}^2 S_x^2 - 2b_0 S_{yx}]. \tag{1.11.1}$$

Theorem1.11.2 The value of b_0 which minimizes $V(\bar{y}_{lr})$ is

$$B = \frac{S_{yx}}{S_x^2} \left(= \rho \frac{S_y S_x}{S_x^2} = \rho \frac{S_y}{S_x} \right)$$ (called the regression linear coefficient of y on x in the population) and the resulting minimum variance is

$$V_{\min}(\bar{y}_{lr}) \frac{1-f}{n} S_y^2 (1-\rho^2). \tag{1.11.2}$$

1.12 Comparison of Linear Regression Estimate with Ratio and mean per unit estimate

For large samples

$$V(\bar{y}_{sr}) = \frac{1-f}{n} S_y^2 \tag{1.12.1}$$

$$V(\bar{y}_R) = \frac{1-f}{n}\left(S_y^2 + R^2 S_x^2 - 2\rho R S_y S_x\right) \tag{1.12.2}$$

$$V(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2 (1-\rho^2) \tag{1.12.3}$$

From equation (1.12.1), and (1.12.3), it is clear that $V(\bar{y}_{lr}) < V(\bar{y}_{sr})$, unless $\rho = 0$, in which case $V(\bar{y}_{lr}) = V(\bar{y}_{sr})$ and the two estimates are equally precise. From equation (1.12.2), and (1.12.3), $\bar{y}_{lr}$ will be precise than $\bar{y}_R$ if and only if $V(\bar{y}_{lr}) < V(\bar{y}_R)$, thus

$$S_y^2(1-\rho^2) < S_y^2 + R^2 S_x^2 - 2\rho R S_y S_x \text{ or}$$

$$-\rho^2 S_y^2 < R^2 S_x^2 - 2\rho R S_y S_x, \text{ or } \rho^2 S_y^2 + R^2 S_x^2 - 2\rho R S_y S_x > 0 \text{ or}$$

$$(\rho S_y - R S_x)^2 > 0), \text{ or } \left(\frac{S_{yx}}{S_y S_x} S_y - R S_x\right)^2 > 0 \text{ or}$$

$$\left(\frac{S_{yx}}{S_x^2} S_x - R S_x\right)^2 > 0, \text{ or } (B-R)^2 S_x^2 > 0 \tag{1.12.4}$$

or $(B-R)^2 > 0$. As the LHS of equation (1.12.4) is a perfect square. Thus we conclude that the linear regression estimate is always better than the ratio estimate expect when $B = R$, i.e. when $B = R$, then

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}) = \bar{y} + \frac{\bar{y}}{\bar{x}}(\bar{X} - \bar{x}) = \bar{y} + \frac{\bar{y}}{\bar{x}}\bar{X} - \bar{y} = \hat{R}\bar{X} = \hat{\bar{Y}}_R.$$

This means that both the estimates linear regression and ratio have the same variance and this occurs only when the regression of y on x is a straight line passes through the origin.

1.13 Regression Estimates in Stratified Sampling

(i) Separate Regression Estimator: A separate regression estimator $\bar{y}_{lrs}$ (s for separate) in stratified sampling may be defined as

$$\bar{y}_{lrs} = \sum_{i=1}^{k} W_{i[\bar{y}_i + b_i(\bar{X}_i - \bar{x}_i)]} = \sum_{i=1}^{k} W_i \bar{y}_{lri} \qquad (1.13.1)$$

where $\bar{y}_{lri} = \bar{y}_i + b_i(\bar{X}_i - \bar{x}_i)$ is the regression estimate for i^{th} stratum mean.

This estimate is appropriate when it is thought that the true regression coefficients B_i vary from stratum to stratum.

Theorem1.13.1 If sampling is independent in different strata and sample size is large enough in each stratum, than $\bar{y}_{lrs}$ is an almost unbiased estimator, and its approximate variance is given by

$$V(\bar{y}_{lrs}) = \sum_{i=1}^{k} \frac{W_i^2(1 - f_i)}{n_i}(S_{yi}^2 + b_i^2 S_{xi}^2 - 2b_i \rho_i S_{yi} S_{xi}) \qquad (1.13.2)$$

Corollary1.13.1 If $b_i = B_i$, the true regression coefficient in stratum i. The minimum value of the variance may be written as

$$V_{\min}(\bar{y}_{lrs}) = \sum_{i=1}^{k} \frac{W_i^2(1 - f_i)}{n_i} S_{yi}^2(1 - \rho_i^2)$$

$$= \sum_{i=1}^{k} \frac{W_i^2(1-f_i)}{n_i}(S_{yi}^2 - \frac{S_{yxi}^2}{S_{xi}^2}) \qquad (1.13.3)$$

(ii) Combined Regression Estimator: A combined regression estimator $\bar{y}_{lrc}$ (c for combined) in stratified sampling may be defined as

$$\bar{y}_{lrc} = \bar{y}_{st} + b_c(\bar{X} - \bar{x}_{st}) \qquad (1.13.4)$$

Where $\bar{y}_{st} = \sum_{i=1}^{k} W_i \bar{y}_i$, $\bar{x}_{st} = \sum_{i=1}^{k} W_i \bar{x}_i$ are the stratified sample means of y and x variates, and b_c is the pooled estimate obtained by

$$b_c = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)(x_{ij} - \bar{x}_i)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

Theorem1.13.2 If sampling is independent in different strata size is large enough in each stratum, the variance of $\bar{y}_{lrc}$ is given by

$$V(\bar{y}_{lrc}) = \sum_{i=1}^{k} \frac{W_i^2(1-f_i)}{n_i}(S_{yi}^2 + b^2 S_{xi}^2 - 2bS_{yxi}) \qquad (1.13.5)$$

Corollary1.13.2 The value of b that minimizes this variance is obtained by minimizing $V(\bar{y}_{lrc})$ with respect to b as

$$\frac{\partial}{\partial b} V(\bar{y}_{lrc}) = \sum_{i=1}^{k} \frac{W_i^2(1-f_i)}{n_i}(2bS_{xi}^2 - 2S_{yxi})$$

$$= \sum_{i=1}^{k} \frac{W_i^2(1-f_i)}{n_i}(bS_{xi}^2 - S_{yxi})$$

$$\Rightarrow b\sum_{i=1}^{k}\frac{W_i^2(1-f_i)}{n_i}S_{xi}^2=\sum_{i=1}^{k}\frac{W_i^2(1-f_i)}{n_i}S_{yxi}, \text{and}$$

$$b=\frac{\sum_{i=1}^{k}\frac{W_i^2(1-f_i)}{n_i}S_{yxi}}{\sum_{i=1}^{k}\frac{W_i^2(1-f_i)}{n_i}S_{xi}^2}=B_c$$

The quantity B_c is a weighted mean of the stratum regression coefficients $B_i=\frac{S_{yxi}}{S_{xi}^2}$.If we write

$$a_i=\frac{W_i^2(1-f_i)}{n_i}S_{xi}^2, \text{ then } B_c=\frac{\sum_{i=1}^{k}a_iB_i}{\sum_{i=1}^{k}a_i}.$$

$$V(\bar{y}_{lrc})=\sum_{i=1}^{k}\frac{W_i^2(1-f_i)}{n_i}(S_{yi}^2+B_c{}^2S_{xi}^2-2B_cS_{yxi}) \qquad (1.13.6)$$

1.14 Bayesians Set-Up

A celebrated result employed somewhere, is Bayes' theorem, named after an English clergyman-Sir Reverend Thomas Bayes. Bayes gave his result in 1763. This fundamental theorem has led to the development of Bayesian theory of statistical Inference, which naturally, finds applications in sampling theory as well. The Bayes' theorem, stated for the discrete case is as follows:

Bayes' Theorem (Discrete Case):

If the k events $B_1,B_2,...,B_k$ are mutually exclusive and exhaustive and A is another event, then the conditional probability

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{i=1}^{k} P(A \mid B_i)P(B_i)} \qquad (1.14.1)$$

is called the Bayes' theorem.

The event A corresponds to experimental outcome and the events B_i to states of the environment. The decision maker is usually given the probabilities $P(A \mid B_i)$ of the experimental outcome A, given the states B_i. He assessed the probabilities $P(B_i)$ of the states B_i in the light of the experimental outcome A .The probability $P(A \mid B_i)$ is termed likelihood which involves the additional information A. The $P(B_i)$ is called the prior probability, and $P(B_i \mid A)$ is called the posterior probability.

Bayes Theorem (Continuous Case)

Suppose that $X' = (X_1, X_2, ..., X_n)$ is a vector of n observations and $P(X \mid \theta)$ be the likelihood of X given θ, where $\theta' = (\theta_1, \theta_2, ..., \theta_k)$ is the vector of k parameters. Suppose also that θ itself has a probability distribution $P(\theta)$. Then the conditional probability $P(\theta \mid X)$ is given by

$$P(\theta \mid X) = \frac{P(X \mid \theta).P(\theta)}{\int_{\theta} P(X \mid \theta).P(\theta)} \qquad (1.14.2)$$

This is known as Bayes' theorem.

$P(\theta)$ is the prior distribution of θ (that is the distribution assigned to θ by the decision maker prior to taking any observations). $P(\theta \mid X)$ is the posterior distribution of θ (that is the prior distribution as revised by the

decision maker through the Bayes theorem in the light of the observation taken). $P(\theta \mid X)$ is the likelihood of X.

Now, we have that

$P(X) = E[P(X \mid \theta)] = \int_{\theta} P(X \mid \theta)P(\theta)d\theta$ a constant (1.15.3) where E denotes the mathematical expectation. In this light Bayes theorem becomes

$$P(\theta \mid X) = C.P(X \mid \theta)P(\theta) \qquad (1.14.3)$$

Thus the denominator in the Bayes theorem is simply a normalizing constant necessary to ensure that the posterior distribution $P(\theta \mid X)$ is a proper density function. That is, it is integrates to one.

In other words, Bayes theorem states that the posterior distribution of θ is proportional to the product of the likelihood and the prior distribution ofθ. That is,

$$P(\theta \mid X) \propto P(X \mid \theta).P(\theta) \qquad (1.14.4)$$

The Bayesian posterior analysis and the Bayesian pre-posterior analysis will now be referred to simply as Bayesian analysis and pre-posterior analysis respectively.

In the Bayesian analysis we base our decision after the experiment has actually or hypothetically been performed and its outcome observed.

In pre-posterior analysis we take decision before performing the experiment actually or hypothetically.

CHAPTER-II

CONSTRUCTION OF STRATA

2.1 Introduction

The basic consideration involved in the construction of strata is that the strata should be internally as homogenous as possible, that is stratum variances S_i^2 are as small as possible. If the distribution of the study variable is available the strata would be created by cutting this distribution at suitable points.

Given the number of strata, Dalenius and Gurney (1951) suggested that the strata boundaries be so determined that $W_i S_i$ remain constant.

Mahalanobis (1952) and Hansen, Hurwitz and Madow (1953) have suggested that strata boundaries be so determined that $W_i S_i$ remain constant. Dalenius and Hodges (1959) have supported the work of Dalenius and Gurney (1951).

Dalenius (1957) has worked out the best stratum boundaries under proportional and Neyman Allocation. Ekman (1959) has suggested approximation to complicated theoretical solutions. Cochran (1961) has examined the applications of these approximations through the empirical studies. Sethi (1963) has showed that the above suggestions fail to provide optimum strata boundaries for certain types of populations. He derived the solutions for optimum stratification points for certain populations. Again, Hess, Sethi and Bal Krishnan (1966) have applied these solutions to some empirical studies and made a comparison of various approximations. Singh and Sukhatme (1969) have suggested several approximate methods to obtain optimal points of stratification.

Singh and Sukhatme (1973) have suggested certain rules for obtaining optimal stratification points based on auxiliary information. Some others who worked out this problem are Singh (1977), Unnithan (1978), Yadav and Singh (1984) etc.

2.2 Fixing the Optimum Stratum Boundaries

The problem of determining the optimum stratum boundaries, when the main study variable is used as stratification variable and a stratified sample, using Neyman allocation (for a fixed total sample size) is adopted to estimate the population mean (or total).

For a variable y the best characteristic is the frequency distribution of y itself. The next best is clearly the frequency distribution of some other quantity highly correlated with y.Given the number of strata, we derive below the equations, for determining the best stratum boundaries, under Neyman allocation. Though, we assume that strata are set up by using the value of y itself.

Dalenius (1950) and Dalenius & Gurney (1951) have developed some rules for the division of the population into strata under proportional and optimum allocation. For the sake of simplicity we shall assume that the population is infinite. Let the $f(y)$ denote the frequency function of the continuous study variable y , $y_0 \le y \le y_k$ where y_0 and y_k are known real numbers and $y_0 < y_k$.

The problem of constructing k strata between y_0 and y_k can then be considered as the problem of determining the $k-1$ stratification points $y_1, y_2, \ldots, y_{k-1}$ such that the sampling variance of the stratified sample mean $\bar{y}_{st}$ is minimum. Where $\bar{y}_{st}$ is the usual estimator of the over all population mean $\bar{Y}$.

Ignoring the finite population correction (f.p.c) the variance of $\bar{y}_{st}$ under Neyman allocation is given as

$$V(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k} W_i S_i\right)^2 \qquad (2.2.1)$$

Where W_i and S_i^2 are the stratum weight (relative frequency) and stratum variance for the $i-$th stratum; $i = 1,2,...,k$ respectively and n is the known fixed total sample size. In order to minimize $V(\bar{y}_{st})$ it is sufficient to minimize $\sum_{i=1}^{k} W_i S_i$ only, because n is a known constant.

As the study variable y is assumed to be continuous, we have

$$W_i = \int_{y_{i-1}}^{y_i} f(t)dt \qquad (2.2.2)$$

$$S_i^2 = \frac{1}{W_i}\int_{y_{i-1}}^{y_i} t^2 f(t)dt - \bar{Y}_i^2 \qquad (2.2.3)$$

Where $\bar{Y}_i = \frac{1}{W_i}\int_{y_{i-1}}^{y_i} t\, f(t)\, dt \qquad (2.2.4)$

is the stratum mean of the i^{th} stratum; $i = 1,2,...,k$.

Let y_0, y_k be the smallest and largest values of y in the population. We have to find intermediate stratum boundaries $y_1, y_2, ..., y_{k-1}$ such that

$$V(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k} W_i S_i\right)^2 \qquad (2.2.1)$$

is a minimum. It is sufficient to minimize, $\sum_{i=1}^{k} W_i S_i$. Thus, since y_i appears in this sum only in the terms $W_i S_i$ and $W_{i+1} S_{i+1}$, we have

$$\frac{\partial}{\partial y_i}(\sum W_i S_i) = \frac{\partial}{\partial y_i}(W_i S_i) + \frac{\partial}{\partial y_i}(W_{i+1} S_{i+1})$$

Let $f(y)$ be the frequency function of y, then

$$W_i = \int_{y_{i-1}}^{y_i} f(t)dt, \qquad \frac{\partial W_i}{\partial y_i} = f(y_i) \tag{2.2.5}$$

Further, since

$$\bar{Y}_i = E(y_i) = \frac{\int_{y_{i-1}}^{y_i} t\, f(t)dt}{\int_{y_{i-1}}^{y_i} f(t)dt}$$

and $S_i^2 = E(y_i^2) - [E(y_i)]^2$

We have

$$W_i S_i^2 = \int_{y_{i-1}}^{y_i} t^2 f(t)dt - \frac{\left(\int_{y_{i-1}}^{y_i} t\, f(t)dt\right)^2}{\int_{y_{i-1}}^{y_i} f(t)dt} \tag{2.2.6}$$

Differentiation of (2.2.6) gives,

$$S_i^2 \frac{\partial W_i}{\partial y_i} + 2W_i S_i \frac{\partial S_i}{\partial y_i} = y_i^2 f(y_i) - 2y_i \bar{Y}_i f(y_i) + \bar{Y}_i^2 f(y_i)$$

Add $S_i^2 \partial W_i / \partial y_i$ to the left side, and the equal quantity $S_i^2 f(y_i)$ to the right side. This gives, on dividing by $2\,S_i$,

$$\frac{\partial(W_i S_i)}{\partial y_i} = S_i \frac{\partial W_i}{\partial y_i} + W_i \frac{\partial S_i}{\partial y_i} = \frac{1}{2} f(y_i) \frac{(y_i - \bar{Y}_i)^2 + S_i^2}{S_i}$$

Similarly we find

$$\frac{\partial(W_{i+1} S_{i+1})}{\partial y_i} = -\frac{1}{2} f(y_i) \frac{(y_i - \bar{Y}_{i+1})^2 + S_{i+1}^2}{S_{i+1}}$$

Hence the calculus equations for y_i are

$$\frac{(y_i - \bar{Y}_i)^2 + S_i^2}{S_i} = \frac{(y_i - \bar{Y}_{i+1})^2 + S_{i+1}^2}{S_{i+1}} \quad (i = 1,2,\ldots,k-1) \qquad (2.2.7)$$

These equations are ill adapted to practical computation, since both $\bar{Y}_i$ and S_i depend on y_i. To come over the difficulty, several quick approximate methods have been provided for, by several research workers on the field. One suggested by Dalenious and Hodges (1959) is given below:

Let

$$Z(y) = \int_{y_0}^{y} \sqrt{f(t)}\,dt$$

If the strata are numerous and narrow, $f(y)$ should be approximately constant (rectangular) within a given stratum. Hence,

$$W_i = \int_{y_{i-1}}^{y_i} f(t)\,dt \approx f_i(y_i - y_{i-1})$$

$$S_i \approx \frac{1}{\sqrt{12}}(y_i - y_{i-1})$$

$$Z_i - Z_{i-1} = \int_{y_{i-1}}^{y_i} \sqrt{f(t)}dt \approx \sqrt{f_i}(y_i - y_{i-1}) = A_i(say) \qquad (2.2.8)$$

Where f_i is the "constant" value of $f(y_i)$ in the stratum i. By putting these approximations, we find

$$\sqrt{12}\sum_{i=1}^{k} W_i S_i \approx \sum_{i=1}^{k} f_i(y_i - y_{i-1})^2 \approx \sum_{i=1}^{k}(Z_i - Z_{i-1})^2 = \sum_{i=1}^{k} A_i^2 \quad (2.2.9)$$

Since $(Z_k - Z_0)$ is fixed, we have that sum on the right hand side of (2.2.9)is minimum when $(Z_i - Z_{i-1})$ that is A_i is a constant.

Thus the rule is:

Given $f(y)$, the rule is to form the cumulative of $\sqrt{f(y)}$ and choose the y_i so that they create equal intervals on the $cum\sqrt{f(y)}$ scale.

From (2.2.9) the rule is equivalent to making $W_i S_i$ approximately constant. But with $W_i S_i$ constant. Neyman allocation gives a constant sample size $n_i = n/k$ in all strata.

Since the optimum is flat with respect to variations in the n_i. Use of the $cum\sqrt{f(y)}$ rule, taking equal sample sizes in the resulting strata, is highly efficient.

Remark: Among the other approximate rules suggested by Ayoma (1954) recommends to make strata of equal width $y_i - y_{i-1}$. Another device by Ekman (1959), is to make the quantity $W_i(y_i - y_{i-1})$ constant.

Cochran (1961) used these rules on a number of actual populations that are skew and found that the rule given by Dalenius and Hodges (1959) worked best.

2.3 The Choice of Strata Boundary on the Basis of Auxiliary Variable when Proportional Allocation is Adopted

The assumption that the stratification is done based on the value of y has only a theoretical aspect but not the practical one, hence unrealistic. In practice some other variable x is used which is correlated with y.Let $f(x,y)$ be the joint probability density function of the variable x and y. If proportional allocation is adopted, then the variance

$$V(\bar{y}_{st})_{prop} = \frac{1}{n}\sum_{i=1}^{k} W_i S_i^2 \quad (f.p.c \text{ ignored}) \text{ is to be minimized.}$$

If $y_0, y_1, \ldots, y_{k-1}$ are the strata boundaries,

$$W_i = \int_{-\infty}^{\infty} \int_{y_{i-1}}^{y_i} dF(x,y) \tag{2.3.1}$$

$$\bar{Y}_i = \frac{1}{W_i} \int_{-\infty}^{\infty} \int_{y_{i-1}}^{y_i} x \; dF(x,y) \tag{2.3.2}$$

$$W_i \; (S_i^2 + \bar{Y}_i^2) = \int_{-\infty}^{\infty} \int_{y_{i-1}}^{y_i} x^2 \; dF(x,y) \tag{2.3.3}$$

We want to minimize $\frac{1}{n}\sum_{i=1}^{k} W_i S_i^2$ with respect to y_i hence differentiating, partially with respect to y_i, y_{i-1} and equating to zero,

we have

$$\frac{d(S_i^2\bar{Y}_i)}{dy_i} + \frac{d(S_{i+1}^2\bar{Y}_{i+1})}{dy_i} = 0 \tag{2.3.4}$$

Let $\int_{-\infty}^{\infty} dF(x,y_i) = \phi(y_i)$ (2.3.5)

$$\frac{dW_i}{dy_i} = \phi(y_i), \quad \frac{dW_{i+1}}{dy_i} = -\phi(y_i) \tag{2.3.6}$$

$$\int_{-\infty}^{\infty} x\, dF(x,y_i) = \phi(y_i)\, E(x \mid y_i) \tag{2.3.7}$$

$$\int_{-\infty}^{\infty} x^2\, dF(x,y_i) = \phi(y_i)\, E(x^2 \mid y_i) \tag{2.3.8}$$

In the light of the above results we have

$$\frac{dW_i\bar{Y}_i}{dy_i} = \phi(y_i)\, E(x \mid y_i),$$

$$\frac{d(W_{i+1}\bar{Y}_{i+1})}{d(y_{i+1})} = -\phi(y_i)\, E(x \mid y_i) \tag{2.3.9}$$

and from (2.3.3) differentiating with respect to y_i we have

$$\frac{dW_i(\bar{Y}_i^2 + S_i^2)}{dy_i} = E(x^2 \mid y_i) \tag{2.3.10}$$

$$\frac{dW_{i+1}(\bar{Y}_i^2 + S_i^2)}{dy_{i+1}} = -E(x^2 \mid y_i) \tag{2.3.11}$$

$$\frac{dW_i S_i^2}{dy_i} = E(x^2 \mid y_i) - \bar{Y}_i^2 \phi(y_i) - 2 W_i \bar{Y}_i \frac{d\bar{Y}_i}{dy_i}$$

$$= E(x^2 \mid y_i) - \bar{Y}_i^2 \phi(y_i) - 2 \bar{Y}_i [\phi(y_i).E(x \mid y_i) - \bar{Y}_i \phi(y_i)] \quad (2.3.12)$$

$$\frac{dW_{i+1} S_{i+1}^2}{dy_i} = -E(x^2 \mid y_i) - Y_{i+1}^2 \phi(y_i) - 2 \bar{Y}_{i+1} [-\phi(y_i).E(x \mid y_i) + \bar{Y}_{i+1} \phi(y_i)] \quad (2.3.13)$$

from (2.3.12) and (2.3.13),we have

$$E(x \mid y_i) = \frac{\bar{Y}_i + \bar{Y}_{i+1}}{2} \quad (2.3.14)$$

the above equation (2.3.14) gives the criteria for choosing the best strata boundaries.

2.4 Approximate Optimum Strata Boundaries for Ratio and Regression Estimator

2.4.1 Preliminaries regarding Optimum Strata Boundaries for Ratio and Regression Estimator

The problem of determining optimum strata boundaries was first considered by Dalenius (1950) and Hayashi et al. (1951). Singh and Sukhatme (1969 and 1973) considered the problem of determining optimum strata boundaries on auxiliary variable x for stratified simple random sampling and ratio and regression estimates respectively. In this paper, we propose an alternative method for determining AOSB for combined ratio and regression estimations. Let the population under consideration be divided into k strata and a stratified simple random

sample of size n be drawn from it, the sample size in the i^{th} strata being n_i so that $\sum_{i=1}^{k} n_i = n$.S

Combined Ratio Estimate

The combined ratio estimate of the population mean Y with X as auxiliary variable is given by (in usual notations)

$$y_{RC} = \left(\sum_{i=1}^{k} W_i y_i \right) X \Big/ \sum_{i=1}^{k} W_i x_i \qquad (2.4.1)$$

and the sample variance of y_{RC} up to the first order of approximation is given by

$$V(y_{RC}) = \sum_{i=1}^{k} n_i^{-1} W_i^2 (\sigma_{iy}^2 - 2R\sigma_{ixy} + R^2\sigma_{ix}^2); \text{R=Y/X} \qquad (2.4.2)$$

Combined Regression Estimate

The combined regression estimate of the population mean Y and its sampling variance (upto the first order of approximation) in usual notations is given by

$$ylc = \sum_{i=1}^{k} W_i y_i + b\left(X - \sum_{i=1}^{k} W_i x_i \right) \qquad (2.4.3)$$

And

$$V(y_{lc}) = \sum_{i=1}^{k} n_i^{-1} W_i^2 (\sigma_{iy}^2 - 2\beta\sigma_{ixy} + \beta^2\sigma_{ix}^2) \qquad (2.4.4)$$

where, $b = \dfrac{1}{\sum_{i=1}^{k} n_i^{-1} W_i^2 S_{ix}^2} \sum_{i=1}^{k} n_i^{-1} W_i^2 S_{ixy}$.

2.4.2 Optimum Allocation

Now we consider allocating the sample optimally to different strata, when the total expected cost C of the survey is fixed. Since the variable x can also be treated as a size measure, we may assume that cost of observing study variable y on a unit is a function of the value of the variable x for that unit. If $c(x)$ is this function then expected cost for observing n_i units in i^{th} stratum is $n_i \mu_{ic}$, where μ_{ic} is the expected value of $c(x)$ in i^{th} stratum (x_{i-1}, x_i) under a marginal density function $f(x)$ of theauxiliary variable x.

The cost function can, therefore, be taken as

$$C = C_0 + \sum_{i=1}^{k} n_i \mu_{ic} + \psi(k) \tag{2.4.5}$$

where C_0 is the overhead cost and $\psi(k)$ is the cost of constructing k strata with $\psi(1) = 0$.

For a given value of k if the variance expressions in (2.4.2) and (2.4.4) are minimized with respect to n_i subject to the condition (2.4.5), the optimum values of n_i for combined ratio reduces to

$$n_i = \frac{[C - C_0 - \psi(K)] Z_{iy} \mu_{ic}}{\sum_{i=1}^{k} W_i^2 \mu_{ic} Z_{iy}^2} \tag{2.4.6}$$

where, $Z_{iy}^2 = \sigma_{iy}^2 - 2R\sigma_{ixy} + R^2 \sigma_{ix}^2$

which when substituted in (2.4.2) gives the minimal variance as

$$V(y_{RC}) = \frac{1}{[C - C_0 - \psi(k)]}\left[\sum_{i=1}^{k} W_i \mu_{ic} Z_{iy}^2\right]^2 \tag{2.4.7}$$

and the optimum values of n_i for combined regression estimate reduces to

$$n_i = \frac{[C - C_0 - \psi(k)] L_{iy} \mu_{ic}}{\sum_{i=1}^{k} W_i^2 \mu_{ic} L_{iy}^2} \tag{2.4.8}$$

where, $L_{iy}^2 = \sigma_{iy}^2 - 2\beta\sigma_{ixy} + \beta^2\sigma_{ix}^2$

The substitution in (2.4.4) gives the minimal variance as

$$V(y_{lc}) = \frac{1}{[C - C_0 - \psi(k)]}\left[\sum_{i=1}^{k} W_i \mu_{ic} L_{iy}^2\right]^2 \tag{2.4.9}$$

The expressions (2.4.7) and (2.4.9) reveal that the two variances are same except the constants R and β. Thus for further discussion, we shall consider only the combined regression estimate with variance given by (2.4.9). The variance in (2.4.9) being a function of strata boundaries on variable x, can further be reduces by using optimum strata boundaries which correspond to minimum of $V(y_{lc})$ with respect to these boundaries.

CHAPTER-III

THE USE OF A RATIO ESTIMATE IN SUCCESSIVE SAMPLING

3.1 Introduction

A task faced annually by the Canadian wildlife services is to obtain reliable estimates of the number of killed birds during the hunting season. Estimates are published for zones within each province. The information is obtained from the National Harvest Survey through a mail questionnaire sent to a stratified random sample of hunters selected from the list of Federal Migratory Game Bird Hunting Permit holders with ecological zones as strata. Because of high variability in kill among hunters it has been a difficult problem to obtain reliable estimates within available resources.

Successive sampling has been discussed in some detail by Jensen [1942], Patterson [1950], Yates [1960], Cochran [1963], Hansen [1953] et al. Sen [1971] and others. Their discussions have, however, been confined to combining a regression estimate from the matched portion of sample with a mean per unit estimate based on current occasion. Because in many surveys with several levels of stratification computations involving regression estimates become relatively complex, we propose to investigate in this paper some theory of successive sampling using a ratio estimate and examine the efficiency of a simpler estimate obtained by giving equal weights to the estimates based on the matched and unmatched portions.

3.2 Theory and Development of Various Methods of Estimating the mean on the second occasion

3.2.1 Selection of the Sample

Let a simple random sample of size n' be selected on the first occasion from a universe of size N and let its mean be $\bar{x}'$. Let a simple random sample of size m be subsampled from the n' units. Let the mean on the first occasion of the sample of size m be $\bar{x}_m$ and on the second occasion $\bar{y}_m$. In addition, let a simple random sample of size u be taken on the second occasion from the universe $N-m$ left after omitting the m units. Let this mean be $\bar{y}_u$ (unmatched portion). Also let the total number of those sampled on the second occasion $(m+u)$ is denoted by n.

3.2.2 Summary of Notation Used

Let $n'(n)$= total sample size on the first (second) occasion.

m = sample size of those questioned on both occasions (matched sample).

$u'(u)$= sample size of those questioned on the first (second) occasion only.

$u' = n' - m$; $u = n - m$

$\bar{x}'(\bar{y}')$=total sample mean on the first (second) occasion estimating $\bar{X}(\bar{Y})$.

$\bar{x}_m(\bar{y}_m)$= matched sample mean first (second) occasion estimating $\bar{X}(\bar{Y})$.

$\hat{R} = \frac{\bar{y}_m}{\bar{x}_m}$ estimating $R = \frac{\bar{Y}}{\bar{X}}$

$\bar{x}_u(\bar{y}_u)$= unmatched sample mean on the first (second) occasion estimating $\bar{X}(\bar{Y})$.

$S_x^2(S_y^2)$= population variance on first (second) occasion.

$\Delta = (S_x / \bar{X})/(S_y / \bar{Y})$, the ratio of the coefficient of variation of the first occasion to that on the second occasion.

ρ= population correlation coefficient.

$Z = \Delta(2\rho - \Delta)$

$\lambda = m/n$

$\theta = n'/n$

Assumptions

(i) Within each year the expectations for the matched and unmatched sample means are equal, i.e, $E(\bar{x}_m) = E(\bar{x}_u)$ and $E(\bar{y}_m) = E(\bar{y}_u)$.Since the samples are random, the sample means will be unbiased estimators of the population mean for each occasion.

(ii) The sampling fractions are small so that the matched and unmatched portions of the sample may be assumed to be independent and the finite population correction factor negligible.

(iii) The standard deviations of $\bar{y}_m$ and $\bar{x}_m$ are small compared to their respective expections. i.e. $\sigma(\bar{x}_m)/E(\bar{x}_m) << 1$ and $\sigma(\bar{y}_m)/E(\bar{y}_m) << 1$

3.3 The Ratio Method of Estimation

a) For an unstratified sample

The matched (m units) and unmatched (u units) portions of the second occasion sample provide independent estimates ($\bar{y}_m$ and $\bar{y}_u$) of the population mean on the second occasion ($\bar{Y}$). For the matched portion an improved estimate, $\bar{y}_m'$ of $\bar{Y}$ may be obtained using a double sampling ratio estimate. Thus

$$\bar{y}_m' = (\bar{y}_m / \bar{x}_m)\bar{x}' = \hat{R}\bar{x}'$$

Applying the argument in section 12.8 of Cochran [1963] we obtain

$$V(\bar{y}_m') = S_y^2 \frac{(n' - u'Z)}{mn'} \qquad (3.3.1)$$

An estimate of the variance may be obtained by replacing S_y^2 and Z in (3.3.1) by their appropriate sample estimates. Since the mean per unit estimate based on m units ($\bar{y}_m$) has variance S_y^2/m, $\bar{y}_m'$ is more efficient than $\bar{y}_m$ when $Z > 0$. When $\Delta = 1$ it is necessary that $\rho > 0.5$ if $\bar{y}_m'$ is to be superior to $\bar{y}_m$ (Cochran [1953], p.165). An estimate, $\bar{y}_r$ of the mean, $\bar{Y}$, of the population on the second occasion is given by combining the two independent estimates, $\bar{y}_m'$ and $\bar{y}_u$ with weights w and $(1-w)$. Thus

$$\bar{y}_r = w\bar{y}_m' + (1-w)\bar{y}_u \qquad (3.3.2)$$

The best estimate of the mean $\bar{Y}$ on the second occasion is obtained by using the values of w in (3.3.2) that would minimize $V(\bar{y}_r)$. Now

$$V(\bar{y}_r) = w^2 V(\bar{y}_m{}') + (1-w)^2 V(\bar{y}_u) \tag{3.3.3}$$

Differentiating (3.3.3) with respect to w and equating to zero gives the value of w which minimizes $V(\bar{y}_r)$ whence

$$\frac{\partial V(\bar{y}_r)}{\partial w} = 2wV(\bar{y}_m{}') - 2(1-w)V(\bar{y}_u) = 0$$

$$wV(\bar{y}_m) = (1-w)V(\bar{y}_u)$$

$$w[V(\bar{y}_m{}') + V(\bar{y}_u)] = V(\bar{y}_u)$$

$$w = \frac{V(\bar{y}_u)}{V(\bar{y}_m{}') + V(\bar{y}_u)} \tag{3.3.4}$$

Also, $V(\bar{y}_u)$ is given by

$$V(\bar{y}_u) = S_y^2 / u \tag{3.3.5}$$

Hence, substituting from (3.3.1) and (3.3.5) into (3.3.4) we obtain

$$w = \frac{S_y^2/u}{\left[\frac{S_y{}^2}{u} + S_y^2\left(\frac{n'-u'Z}{mn'}\right)\right]}$$

$$= \frac{mn'S_y^2}{(mn' + n'u - uu'z)S_y^2}$$

Put $u = n - m$ and $u' = n' - m$

$$= \frac{mn'}{mn' + n'(n-m) - (n-m)(n'-m)Z}$$

$$w=\frac{n^2\left(\frac{mn'}{n^2}\right)}{n^2\left[\frac{nn'-Z(nn'-mn'-mn+mm)}{n^2}\right]}$$

Put $\lambda=m/n$ and $\theta=n'/n$

$$w=\frac{\theta\lambda}{\theta-Z(\theta-\lambda)(1-\lambda)} \tag{3.3.6}$$

Again substituting from (3.3.6), (3.3.5) and (3.3.1) into (3.3.3) we have

$$V(\bar{y}_r)=\left[\frac{\theta^2\lambda^2S_y^2(n'-u'Z)}{mn'\{\theta-Z(\theta-\lambda)(1-\lambda)\}^2}+\frac{S_y^2}{u}\left(1-\frac{\theta\lambda}{\theta-Z(\theta-\lambda)(1-\lambda)}\right)^2\right]$$

$$=\frac{1}{[\theta-Z(\theta-\lambda)(1-\lambda)]^2}\left[\frac{\theta^2\lambda^2S_y^2(n'-u'Z)}{mn'}+\frac{S_y^2}{u}\{\theta-Z(\theta-\lambda)(1-\lambda)\}^2\right]$$

$$=\frac{S_y^2}{[\theta-Z(\vartheta-\lambda)(1-\lambda)]^2}\left[\frac{\theta^2\lambda^2(n'-u'Z)}{mn'}+\frac{1}{u}\{\theta(1-\lambda)-Z(\theta-\lambda)(1-\lambda)\}^2\right]$$

Put $\theta=n'/n$ and $\lambda=m/n$

$$=\frac{S_y^2n^4}{[nn'-Z(n'-m)(n-m)]^2}\frac{1}{n^4}\left[mn'(n'-u'Z)+\frac{1}{u}(n-m)^2\{n'-Z(n'-m)\}^2\right]$$

Put $u=n-m$ and $u'=n'-m$

$$=\frac{S_y^2}{[nn'-uu'Z]^2}[mn'(n'-u'Z)+u(n'-u'Z)^2]$$

$$=\frac{S_y^2}{[nn'-uu'Z]^2}(n'-u'Z)[mn'+un'-uu'Z]$$

$$=\frac{S_y^2}{[nn'-uu'Z]^2}(n'-u'Z)[n'(m+u)-uu'Z]$$

$$=\frac{S_y^2}{[nn'-uu'Z]^2}(n'-u'Z)[n'(m+n-m)-uu'Z]$$

$$=\frac{S_y^2}{[nn'-uu'Z]^2}(n'-u'Z)[nn'-uu'Z]$$

$$V(\bar{y}_r)=\frac{S_y^2(n'-u'Z)}{(nn'-uu'Z)} \tag{3.3.7}$$

If, however, the estimate $\bar{y}'$ of the mean $\bar{Y}$ on the current occasion were based exclusively on the n sampling units for the second occasion its variance would be

$$V(\bar{y}')=\frac{S_y^2}{n} \tag{3.3.8}$$

The gain in precision, G_1 of $\bar{y}_r$ over $\bar{y}'$ is given by

$$G_1=\{V(\bar{y}')-V(\bar{y})\}/V(\bar{y}_r)$$

$$G_1=\frac{\frac{S_y^2}{n}-S_y^2\left(\frac{n'-u'Z}{nn'-uu'Z}\right)}{S_y^2\left(\frac{n'-u'Z}{nn'-uu'Z}\right)}$$

$$=\frac{nn'-uu'Z-nn'+nu'Z}{nn'-nu'Z}$$

$$= \frac{Z(un' - uu')}{nn' - nu'Z}$$

Put $u = n - m$ and $u' = n' - m$

$$= \frac{Z[n(n' - m) - (n' - m)(n - m)]}{nn' - nZ(n' - m)}$$

$$= \frac{Z\left[n^2\left(\frac{n' - m}{n}\right) - n^2\left(\frac{n' - m}{n}\right)\left(\frac{n - m}{n}\right)\right]}{n^2\left[\frac{n'}{n} - Z\left(\frac{n' - m}{n}\right)\right]}$$

Put $\theta = n'/n$ and $\lambda = m/n$

$$= \frac{Z[(\theta - \lambda) - (\theta - \lambda)(1 - \lambda)]}{\theta - Z(\theta - \lambda)}$$

$$= \frac{Z(\theta - \lambda)[1 - 1 + \lambda]}{\theta - Z(\theta - \lambda)}$$

$$G_1 = \frac{Z\lambda(\theta - \lambda)}{\theta - Z(\theta - \lambda)} \qquad (3.3.9)$$

Necessarily $\lambda \leq \min(\theta, 1)$. If $\lambda = \theta$ or $\lambda = 0$, the gain is zero. For other λ there will be positive gain if $\rho > \Delta/2$ and loss if $\rho < \Delta/2$, as has been noted by Cochran [1963], i.e. $Z > 0$ will give positive gain. G_1 increases with increasing θ and increasing ρ and as Δ approaches ρ.

Differentiating G_1 with respect to λ and equating to zero we obtain

$$\frac{\partial G_1}{\partial \lambda} = [\{\theta - Z(\theta - \lambda)\}(\theta - 2\lambda) - Z\lambda(\theta - \lambda)] = 0$$

$$[\theta - Z(\theta - \lambda)](\theta - 2\lambda) = Z\lambda(\theta - \lambda)$$

$$\theta - Z(\theta - \lambda) = \frac{Z\lambda(\theta - \lambda)}{\theta - 2\lambda}$$

$$\theta = \frac{Z\lambda(\theta - \lambda)}{\theta - 2\lambda} + Z(\theta - \lambda)$$

$$\theta = \frac{Z\lambda(\theta - \lambda) + Z(\theta - \lambda)(\theta - 2\lambda)}{(\theta - 2\lambda)}$$

$$\theta = \frac{Z(\theta - \lambda)[\lambda + \theta - 2\lambda]}{(\theta - 2\lambda)}$$

$$\theta = \frac{Z(\theta - \lambda)^2}{(\theta - 2\lambda)}$$

$$\theta^2 - 2\theta\lambda = Z(\theta - \lambda)^2$$

$$\theta^2 - 2\theta\lambda = Z\theta^2 + Z\lambda^2 - 2Z\theta\lambda$$

$$\theta^2 - 2\theta\lambda - Z\theta^2 - Z\lambda^2 + 2Z\theta\lambda = 0$$

Add and subtract λ^2 we obtain

$$\theta^2 - 2\theta\lambda - Z\theta^2 - Z\lambda^2 + 2Z\theta\lambda - \lambda^2 + \lambda^2 = 0$$

$$\theta^2(1 - Z) + \lambda^2(1 - Z) - 2\theta\lambda(1 - Z) - \lambda^2 = 0$$

$$(1 - Z)(\theta - \lambda)^2 = \lambda^2$$

$$(1 - Z) = \frac{\lambda^2}{(\theta - \lambda)^2}$$

$$\sqrt{1 - Z} = \frac{\lambda}{\theta - \lambda}$$

$$(\theta - \lambda)\sqrt{1-Z} = \lambda$$

$$\theta\sqrt{1-Z} - \lambda\sqrt{1-Z} = \lambda$$

$$\theta\sqrt{1-Z} = \lambda + \lambda\sqrt{1-Z}$$

$$\theta\sqrt{1-Z} = \lambda(1+\sqrt{1-Z})$$

$$\lambda = \frac{\theta\sqrt{1-Z}}{1+\sqrt{1-Z}}$$

Whence

$$\lambda_{opt} = \min\left\{\frac{\theta\sqrt{1-Z}}{1+\sqrt{1-Z}}, 1\right\} \quad \text{if } Z > 0$$

$$0 \text{ or } \theta \quad \text{if } Z < 0$$

Since $Z > 0$ implies $d^2G_1/d\lambda^2 < 0$, but $Z < 0$ implies $d^2G_1/d\lambda^2 > 0$ and the best value of λ is an extreme. Finally, when $Z = 0$, G_1 is always 0 and all values of λ are optimal.

Estimates for quantities such as $\lambda_{opt}, w, G_1, V(\bar{y}_r)$, etc., can be obtained by replacing the population values on the right sides of the expressions by sample estimates from the appropriate year's sample. In the case of λ_{opt} one requires some idea of these values before selecting the sample and estimates of S_y^2, $\bar{Y}$ and hence R must be made from the previous year's sample (i.e. assume $\Delta = 1$). ρ is harder to determine, but similar data or data from two consecutive previous years would provide a guide.

When $\lambda = \lambda_{opt}$, the variance of $\bar{y}_r$ is given by

From (3.3.7)

$$V(\bar{y}) = \frac{S_y^2[n' - u'Z]}{nn' - uu'Z}$$

Put $u = n - m$ and $u' = n' - m$

$$= \frac{S_y^2[n' - Z(n' - m)]}{nn' - Z(n-m)(n'-m)}$$

$$= \frac{nS_y^2\left[\frac{n'}{n} - Z\left(\frac{n'-m}{n}\right)\right]}{n^2\left[\frac{n'}{n} - Z\left(\frac{n-m}{n}\right)\left(\frac{n'-m}{n}\right)\right]}$$

Put $\theta = n'/n$ and $\lambda = m/n$

$$= \frac{S_y^2[\theta - Z(\theta - \lambda)]}{n[\theta - Z(1-\lambda)(\theta-\lambda)]} \qquad (3.3.11a)$$

From (3.3.11), we obtain

$$= \frac{S_y^2\left[\theta - Z\left(\theta - \frac{\theta\sqrt{1-Z}}{1+\sqrt{1-Z}}\right)\right]}{n\left[\theta - Z\left(1 - \frac{\theta\sqrt{1-Z}}{1+\sqrt{1-Z}}\right)\left(\theta - \frac{\theta\sqrt{1-Z}}{1+\sqrt{1-Z}}\right)\right]}$$

$$= \frac{S_y^2\left[\frac{\theta + \theta\sqrt{1-Z} - Z\theta - Z\theta\sqrt{1-Z} + Z\theta\sqrt{1-Z}}{1+\sqrt{1-Z}}\right]}{n\left[\theta - Z\left(\frac{1+\sqrt{1-Z} - \theta\sqrt{1-Z}}{1+\sqrt{1-Z}}\right)\left(\frac{\theta + \theta\sqrt{1-Z} - \theta\sqrt{1-Z}}{1+\sqrt{1-Z}}\right)\right]}$$

$$=\frac{S_y^2(\theta+\theta\sqrt{1-Z}-Z\theta)(1+\sqrt{1-Z})}{n[\theta+\theta(1-Z)+2\theta\sqrt{1-Z}-Z\theta(1+\sqrt{1-Z}-\theta\sqrt{1-Z})]}$$

$$V(\bar{y}_{r(opt)})=\frac{S_y^2[2\theta(1-Z)+2\theta\sqrt{1-Z}-Z\theta\sqrt{1-Z}]}{n[2\theta(1-Z)+2\theta\sqrt{1-Z}-Z\theta\sqrt{1-Z}+Z\theta^2\sqrt{1-Z}]}$$

From (3.3.11a), we get

$$V(\bar{y})=\frac{S_y^2[\theta-Z(\theta-\lambda)]}{n[\theta-Z(1-\lambda)(\theta-\lambda)]}$$

Put $\lambda_{opt=1}$, we obtain

$$V(\bar{y}_{r(opt)})=\frac{S_y^2[\theta-Z(\theta-1)]}{n[\theta-(1-1)(\theta-1)]}$$

$$=\frac{S_y^2(\theta-Z\theta+Z)}{n\theta}$$

$$=\frac{S_y^2}{n}\left(1-Z+\frac{Z}{\theta}\right)$$

$$V(\bar{y}_{r(opt)})=\frac{S_y^2}{n}\frac{[2\theta(1-Z)+2\theta\sqrt{1-Z}-Z\theta\sqrt{1-Z}]}{[2\theta(1-Z)+2\theta\sqrt{1-Z}-Z\theta\sqrt{1-Z}+Z\theta^2\sqrt{1-Z}]}$$

if $\lambda_{opt}<1$

$$\frac{S_y^2}{n}\left(1-Z-\frac{Z}{\theta}\right) \text{ if} \lambda_{opt=1}$$

The gain in precision over a mean per unit estimate when $\lambda=\lambda_{opt}$ is

When $\lambda_{opt}<1$

$$G_{1(\max)}=\frac{V(\bar{y}')-V(\bar{y}_{r(opt)})}{V(\bar{y}_{r(opt)})}$$

$$=\frac{\frac{S_y^2}{n}-\frac{S_y^2}{n}\frac{[2\theta(1-Z)+2\theta\sqrt{1-Z}-Z\theta\sqrt{1-Z}]}{[2\theta(1-Z)+2\theta\sqrt{1-Z}-Z\theta\sqrt{1-Z}+Z\theta^2\sqrt{1-Z}]}}{\frac{S_y^2}{n}\left[\frac{2\theta(1-Z)+2\theta\sqrt{1-Z}-Z\theta\sqrt{1-Z}}{2\theta(1-Z)+2\theta\sqrt{1-Z}-Z\theta\sqrt{1-Z}+Z\theta^2\sqrt{1-Z}}\right]}$$

$$=\frac{2\theta(1-Z)+2\theta\sqrt{1-Z}-Z\theta\sqrt{1-Z}+Z\theta^2\sqrt{1-Z}-2\theta(1-Z)}{2\theta(1-Z)+2\theta\sqrt{1-Z}-Z\theta\sqrt{(1-Z)}}$$

$$-\frac{2\theta\sqrt{1-Z}-Z\theta\sqrt{1-Z}}{2\theta(1-Z)+2\theta\sqrt{1-Z}-Z\theta\sqrt{(1-Z)}}$$

$$=\frac{Z\theta^2\sqrt{1-Z}}{2\theta(1-Z)+2\theta\sqrt{1-Z}-Z\theta\sqrt{1-Z}}$$

$$=\frac{Z\theta^2\sqrt{1-Z}}{\theta\sqrt{1-Z}(2\sqrt{1-Z}+2-Z)}$$

$$=\frac{Z\theta}{2\sqrt{1-Z}+1+(1-Z)}$$

$$=\frac{Z\theta}{(1+\sqrt{1-Z})^2}$$

when $\lambda_{opt=1}$, we obtain

$$G_{1(\max)}=\frac{V(\bar{y}')-V(\bar{y}_{r(opt)})}{V(\bar{y}_{r(opt)})}$$

$$=\frac{\frac{S_y^2}{n}-\frac{S_y^2}{n}\left(1-Z+\frac{Z}{\theta}\right)}{\frac{S_y^2}{n}\left(1-Z+\frac{Z}{\theta}\right)}$$

$$= \frac{\theta - \theta + Z\theta - Z}{\theta - Z\theta + Z}$$

$$= \frac{Z(\theta - 1)}{\theta - Z(\theta - 1)}$$

$$G_{1(\max)} = \frac{Z\theta}{(1+\sqrt{1-Z})^2} \qquad \text{if } \lambda_{opt<1}$$

$$\frac{Z(\theta - 1)}{\theta - Z(\theta - 1)} \qquad \text{if } \lambda_{opt=1}$$

b) For a stratified sample:

In a stratified random design the estimator of the overall mean is given by

$$\bar{y}_{st} = \sum_i \alpha_i \bar{y}_i \qquad (3.3.12)$$

Where α_i is the weight of the i^{th} stratum and $\bar{y}_i$ is the sample mean of the i^{th} stratum.

If a double sampling design is applied then the $\bar{y}_i$ in (3.3.12) are replaced by double sampling estimators of the mean, $\bar{y}_{ri}$ and thus the overall mean becomes

$$\bar{y}_{rst} = \sum_i \alpha_i \bar{y}_{ri} \qquad (3.3.13)$$

It may happen that double sampling is not possible in some strata. An example of this is the National Harvest Survey where there are two obvious strata- (1) Those hunters who did not buy a Federal Migratory Game Bird Hunting Permit in the previous year (the first occasion) and were thus non-hunters, and (2) Those who did buy. Clearly, group 2 leads itself to double sampling whereas group 1 does not since it was not part

of the universe on the first occasion. Hence, (3.3.13) will be a combination of double and single sample means.

To determine the best sample allocation for a given total sample size

$$V(\bar{y}_{rst}) = \sum_i \alpha_i^2 V(\bar{y}_{ri})$$

can be minimized and hence the sample sizes for each stratum can be determined. In practice this is a very laborious procedure and the accuracy with which the necessary parameters can be estimated may not justify this.

Alternatively, sample sizes within strata can be determined by minimizing $V(\bar{y}_{st})$, i.e. allocating the sample between strata so as to obtain the optimum stratified random sample estimator. Then, within each stratum, a portion of the sample could be matched using the estimated λ_{opt}. In this case, the gain of $\bar{y}_{rst}$ over $\bar{y}_{st}$ is

$$G_{st} = \frac{V(\bar{y}_{st}) - V(\bar{y}_{rst})}{V(\bar{y}_{rst})}$$

$$= \frac{\sum_i \alpha_i^2 V(\bar{y}_i) - \sum_i \alpha_i^2 V(\bar{y}_{ri})}{\sum_i \alpha_i^2 V(\bar{y}_{ri})}$$

$$= \frac{\sum_i \alpha_i^2 \left(\frac{V(\bar{y}_i) - V(\bar{y}_{ri})}{V(\bar{y}_{ri})} \right) V(\bar{y}_{ri})}{\sum_i \alpha_i^2 V(\bar{y}_{ri})}$$

$$= \frac{\sum_i \alpha_i^2 V(\bar{y}_{ri}) G_i}{\sum_i \alpha_i^2 V(\bar{y}_{ri})}$$

Where $G_i = [V(\bar{y}_i) - V(\bar{y}_{ri})]/V(\bar{y}_{ri})$ is the gain in the i^{th} stratum. If the i^{th} stratum was not amenable to double sampling, then $\bar{y}_{ri} = \bar{y}_i$ and hence $G_i = 0$.

3.4 **Comparison of Ratio and Regression Estimates**

An estimate which has been more fully discussed in the literature is the regression estimate given by (3.1) of Patterson [1950]. A regression estimate of the mean from the matched portion is combined with the mean from the unmatched portion with weights chosen to minimize its variance$(\bar{y}_{reg})$.

$$\bar{y}_{reg} = \frac{t\{\bar{y}_m + (\rho S_y / S_x)(\bar{x}' - \bar{x}_m)\} + u\bar{y}_u}{t+u}$$

Where

$$\frac{1}{t} = \frac{\rho^2}{n'} + \frac{1-\rho^2}{m}$$

The variance of $\bar{y}_{reg}$ is then

$$V(\bar{y}_{reg}) = \frac{S_y^2(n' - \rho^2 u')}{n'n - \rho^2 u'u}$$

The gain in precision of regression estimate over the ratio estimate is

$$G_2 = \frac{\lambda\theta(\theta-\lambda)(\rho^2 - Z)}{\{\theta - Z(\theta-\lambda)(1-\lambda)\}\{\theta - \rho^2(\theta-\lambda)\}}$$

G_2 is generally small expect where Δ is large. For fixed λ the advantage of using a regression estimate decreases to zero as ρ approaches Δ. This

is also true if the regression estimate using its optimum matched portion (λ) is compared to the ratio estimate using its optimum matched portion. Generally, the optimum portion to match for the regression estimate is not the same as that for the ratio estimate.

CHAPTER-IV

DOUBLE SAMPLING

4.1 Double Sampling for Ratio Estimator:

In many situations, the population mean of auxiliary variable $\bar{X}$ is not known and the ratio estimator can not be used to estimate the population mean $\bar{Y}$ of study variable. In such situations, the procedure is to use the method of double sampling, which consists of drawing a first sample of size n' by simple random sampling and only auxiliary variable is measured to estimate the population mean $\bar{X}$ and then a sub-sample of size n from n' is drawn and both auxiliary and main variables are measured to estimate the mean of the under study variable. The ratio estimate of the population mean $\bar{Y}$ in double sampling is defined as

$$\bar{y}_{Rd} = \frac{\bar{y}}{\bar{x}}\bar{x}' = \hat{R}\bar{x}'$$

where $\bar{y}$, and $\bar{x}$ are sub-sample means of main and auxiliary variables respectively, and $\bar{x}'$ is the mean of auxiliary variable in the first sample.

To find the approximate variance, write

$$\bar{y}_{Rd} - \bar{Y} = \frac{\bar{y}}{\bar{x}}\bar{x}' - \bar{Y} = \frac{\bar{y}}{\bar{x}}\bar{x}' - \frac{\bar{y}}{\bar{x}}\bar{X} + \frac{\bar{y}}{\bar{x}}\bar{X} - \bar{Y}$$

$$= \left(\frac{\bar{y}}{\bar{x}}\bar{X} - \bar{Y}\right) + \frac{\bar{y}}{\bar{x}}(\bar{x}' - \bar{X}) = \left(\frac{\bar{y}}{\bar{x}}\bar{X} - R\bar{X}\right) + \frac{\bar{y}}{\bar{x}}(\bar{x}' - \bar{X})$$

$$= \frac{\bar{X}}{\bar{x}}(\bar{y} - R\bar{x}) + \frac{\bar{y}}{\bar{x}}(\bar{x}' - \bar{X})$$

Suppose n is sufficiently large, so that $\frac{\bar{X}}{\bar{x}} \cong 1$, and $\frac{\bar{y}}{\bar{x}} \cong R$, then

$$\bar{y}_{Rd} - \bar{Y} = (\bar{y} - R\bar{x}) + R(\bar{x}' - \bar{X}) \tag{4.1.1}$$

Two cases may arise:

Case1: The first and second samples are independent, then the terms in the RHS of equation (4.1.1) are also independent, and we have

$$V[\bar{y}_{Rd} - \bar{Y}] = V(\bar{y}_{Rd}) = V(\bar{y} - R\bar{x}) + R^2 V(\bar{x}' - \bar{X})$$

$$= V(\bar{y}) + R^2 V(\bar{x}) - 2RCov(\bar{x}, \bar{y}) + R^2 V(\bar{x}')$$

$$= \left(\frac{N-n}{nN}\right) S_y^2 + R^2 \left(\frac{N-n}{nN}\right) S_x^2 - 2R \left(\frac{N-n}{nN}\right) S_{yx} + R^2 \left(\frac{N-n}{nN}\right) S_x^2$$

After ignoring fpc, $V(\bar{y}_{Rd})$ reduces to

$$V(\bar{y}_{Rd}) = \frac{S_y^2}{n} + R^2 \frac{S_x^2}{n} - 2R \frac{S_{yx}}{n} + R^2 \frac{S_x^2}{n'}$$

$$= \frac{S_y^2 + R^2 S_x^2 - 2RS_{yx}}{n} + \frac{R^2 S_x^2}{n'}$$

$$= \frac{V}{n} + \frac{V'}{n'}$$

Where, $V = S_y^2 + R^2 S_x^2 - 2RS_{yx}$ and $V' = R^2 S_x^2$

Case 2: When the second sample is a random sub-sample of the first sample, and we have

$$V[\bar{y}_{Rd} - \bar{Y}] = V(\bar{y}_{Rd}) = V_1[E_2(\bar{y}_{Rd} - \bar{Y})] + E_1[V_2(\bar{y}_{Rd} - \bar{Y})]$$

Consider

$$E_2(\bar{y}_{Rd} - \bar{Y}) = E_2(\bar{y} - R\bar{x}) + RE_2(\bar{x}' - \bar{X}) = \bar{y}' - R\bar{x}' = \bar{y}' - \bar{Y}$$, for large n'.

And

$$V_2(\bar{y}_{Rd} - \bar{Y}) = V_2(\bar{y} - R\bar{x}) + R^2 V_2(\bar{x}' - \bar{X}) = V_2(\bar{y} - R\bar{x})$$

Define a variate

$$d_i = y_i - Rx_i, \qquad i = 1,2,...,n'$$

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - Rx_i) = \bar{y} - R\bar{x} \qquad \text{and}$$

$$V(\bar{d}) = \left(\frac{1}{n} - \frac{1}{n'}\right)S_d'^2 = V_2(\bar{y} - R\bar{x})$$

Therefore,

$$V(\bar{y}_{Rd}) = V_1(\bar{y}' - \bar{Y}) + E_1\left[\left(\frac{1}{n} - \frac{1}{n'}\right)S_d'^2\right]$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)E_1(S_d'^2)$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_d^2$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)(S_y^2 + R^2S_x^2 - 2RS_{yx})$$

$$= \frac{S_y^2}{n'} + \frac{S_y^2 + R^2S_x^2 - 2RS_{yx}}{n} - \frac{S_y^2 + R^2S_x^2 - 2RS_{yx}}{n'}$$

$$= \frac{S_y^2 + R^2 S_x^2 - 2RS_{yx}}{n} + \frac{2RS_{yx} - R^2 S_x^2}{n'}$$

$$= \frac{V}{n} + \frac{V'}{n'}$$

Where, $V = S_y^2 + R^2 S_x^2 - 2RS_{yx}$, and $V' = 2RS_{yx} - R^2 S_x^2$

4.2 Double Sampling for Regression Estimator:

Suppose, the population is infinite and relation between y_i and x_i is linear giving by the model

$$y_i = \bar{Y} + B(x_i - \bar{X}) + e_{i\alpha} \tag{4.2.1}$$

Where $e_{i\alpha}$ is a random variable with $E(e_{i\alpha} / x_i) = 0$, and

$$V(e_{i\alpha}) = E(e_{i\alpha} / x_i)^2 = S_e^2 = \frac{1}{N-1} \sum e_{i\alpha}^2$$

Let first large sample of size n' is drawn and only auxiliary variable x_i is measured to estimate $\bar{X}$. In the second phase, a sample of size n is drawn and both x_i and y_i are measured.

Regression estimate of population mean $\bar{Y}$ in double sampling is defined by

$$\bar{y}_{lrd} = \bar{y} + b(\bar{x}' - \bar{x}) \tag{4.2.2}$$

Where, b is the least square estimate of B given by

$$b = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Averaging equation (4.2.1) over the sample

$$\bar{y} = \bar{Y} + B(\bar{x} - \bar{X}) + \bar{e} \tag{4.2.3}$$

$$E(\bar{e}) = 0,\ V(\bar{e}) = \frac{1}{n} S_e^2$$

From equation (4.2.1) and (4.2.3), we have

$y_i - \bar{y} = B(x_i - \bar{x}) + e_{i\alpha} - \bar{e}$, substituting this in b

$$b = \frac{\sum \{B(x_i - \bar{x}) + (e_{i\alpha} - \bar{e})\}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$= \frac{B\sum (x_i - \bar{x})^2 + \sum (e_{i\alpha} - \bar{e})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$= B + \frac{\sum_i (e_{i\alpha} - \bar{e})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

$$= B + \frac{\sum_i e_{i\alpha}(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}, \text{ as } \sum_i \bar{e}(x_i - \bar{x}) = 0 \tag{4.2.4}$$

From relation (4.2.2)

$$\bar{y}_{lrd} - \bar{Y} = B(\bar{x} - \bar{X}) + \bar{e} + \left[B + \frac{\sum e_{i\alpha}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right] (\bar{x}' - \bar{x})$$

$$= B(\bar{x} - \bar{X}) + \bar{e} + \left[\frac{\sum e_{i\alpha}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right] + B(\bar{x}' - \bar{x})$$

$$= \bar{e} + \left[\frac{(\bar{x}' - \bar{x}) \sum e_{i\alpha} (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right] + B(\bar{x}' - \bar{X}) \qquad (4.2.5)$$

Now, using the technique of regression theory, taking expectation of both sides keeping the value x_i fixed in both the samples

$E(\bar{y}_{lrd} - \bar{Y}) = B(\bar{x}' - \bar{X})$, as $E(\bar{e}) = 0$, $E(e_{i\alpha} / x_i) = 0$, from (4.2.1)

$\neq 0$

Therefore, under the condition of that $x_i's$ are fixed $\bar{y}_{lrd}$ is bias estimate of $\bar{Y}$.

$$V_e(\bar{y}_{lrd}) = V_e(\bar{y}_{lrd} - \bar{Y}) + (bias)^2$$

$$= V(\bar{e}) + (\bar{x}' - \bar{x})^2 \frac{V(e_{i\alpha}) \sum (x_i - \bar{x})^2}{\left[\sum_i (x_i - \bar{x}) \right]^2} + B^2 (\bar{x}' - \bar{X})^2$$

$$= \frac{S_e^2}{n} + \frac{S_e^2 (\bar{x}' - \bar{x})^2}{\sum (x_i - \bar{x})^2} + B^2 (\bar{x}' - \bar{X})^2$$

In regression theory $S_e^2 = S_y^2 (1 - \rho^2)$

$$V_e(\bar{y}_{lrd}) = S_y^2 (1 - \rho^2) \left[\frac{1}{n} + \frac{(\bar{x}' - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] + B(\bar{x}' - \bar{X})^2$$

To find $V(\bar{y}_{lrd})$ i.e., variance of $\bar{y}_{lrd}$ overall possible values of $x_i's$ in two samples.

$$V(\bar{y}_{lrd}) = E[V_e(\bar{y}_{lrd})] = S_y^2 (1 - \rho^2) \left[\frac{1}{n} + \frac{(\bar{x}' - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] + B^2 E(\bar{x}' - \bar{X})^2$$

4.3 Double Sampling for Ratio and Regression Estimation with Sub Sampling of Non-Respondents

4.3.1 Preliminaries regarding Double Sampling for Ratio and Regression Estimation with Sub Sampling of Non-Respondents

In many human surveys, information is in most cases not obtained from all the units in the survey even after some call-backs. An estimate obtained from such incomplete data may be misleading especially when the respondents differ from the non-respondents because the estimate can be biased. Hansen and Hurwitz (1946) proposed a technique for adjusting for non-response to address the bias problem. Their idea is to take a sub-sample from the non-respondents to get an estimate for the subpopulation represented by the non-respondents.

Cochran (1977), using Hansen and Hurwitz (1946) procedure, proposed the ratio and regression estimators of the population mean of the study variable in which information on the auxiliary variable is obtained from all the sample units, while some sample units failed to supply information on the study variable. In addition, the population mean of the auxiliary variable is known. Here we assume that the population mean of the auxiliary variable is not known. We, therefore, use the double sampling method to estimate the mean of the auxiliary variable and then go on to estimate the mean of the study variable in a similar manner as Cochran (1977).

In practice, non-response is often compensated for by weighting adjustment (Oh and Scheuren 1983) or by imputation (Kalton and Karsprzyk 1986). The procedures used for weighting adjustment and imputation strive for elimination of the bias due to non-response.

However, those procedures are based on untenable assumptions on the response mechanism. When the assumed mechanism is wrong, then the resulting estimate can be seriously biased. Moreover, it is difficult to eliminate the bias entirely when non-response is confounded in the sense that the response probability is dependent on the survey character. Rancourt, Lee, and Sarndal (1994) provided a partial correction for the situation. Hansen and Hurwitz's sub-sampling approach does not have this defect although it costs more because of extra work required for sub-sampling the non-respondents. Nonetheless, if the bias problem is serious, the procedure is viable options to address the problem without resorting to 100 percent response, which can be very expensive.

In the next section, double sampling ratio and regression estimators are considered. Generally, the double sampling procedure is used when it is necessary to make use of auxiliary information to improve the precision of an estimate but the population distribution of the auxiliary information is not known. The first phase sample is used to estimate the population distribution of the auxiliary variable, while the second phase sample is used to obtain the required information on the variable of main interest. The optimum sampling fractions are derived for the estimators for a fixed cost. The performances of the proposed estimators are compared both theoretically and empirically with the Hansen and Hurwitz estimator.

4.3.2 The Double Sampling for Ratio and Regression Estimators in the presence of Non-response

4.3.2.1 Background

To estimate the population mean $\overline{X}$ of the auxiliary variable, a large first phase sample of size n' is selected from N units in the population by simple random sampling without replacement (SRSWOR). A smaller

second phase sample of size n is selected from n' by SRSWOR and the character y is measured on it. The ratio estimator of the mean of y is $\bar{y}'_r = (\bar{y}/\bar{x})\bar{x}'$, where $\bar{x}'$ is the sample mean from n' units, $\bar{y}$ and $\bar{x}$ are obtained from the second phase sample if there is no non-response in the second phase sample. If, however, there is non-response in the second phase sample, we may use an estimator obtained from only the respondents or take a sub-sample of the non-respondents and re-contact them. The former option is much cheaper than the latter because securing missing information from the non-respondents by re-contact requires usually much more effort and cost. However, it is quite feasible that the non-respondents differ significantly in the main character from the respondents so that a serious bias results. In this situation, sub-sampling of the non-respondents may be beneficial. Hence, we pursue the sub-sampling idea of Hansen and Hurwitz for a double sampling situation. Basically, the estimators proposed here are double sampling version of Cochran (1977, page 374), that is, double sampling ratio and regression estimators for $\bar{Y}$ adjusted for non-response by using the Hansen and Hurwitz (1946) procedure.

Let us assume that all the n' units supplied information on the auxiliary variable x at the first phase. But let n_1 units supply information on y and n_2 refuse to respond at the second phase. From the n_2 non-respondents, an SRSWOR of m units is selected with the inverse sampling rate k, where $m = n_2/k$, $k > 1$. All the m units respond this time around. This can be applied in a household survey where the household size is used as an auxiliary variable for the estimation of, say, family expenditure. Information can be obtained completely on the family size during the household listing while there may be non-response on the household expenditure.

In the following presentation, we assume that the whole population (denoted by A) is satisfied into two strata: one is the stratum (denoted by A_1) of N_1 units, which would respond on the first call at the second phase and the other stratum (denoted by A_2) consists of N_2 units which would not respond on the first call at the second phase but will respond on the second call. Let the first and second phase samples be denoted by a' and a respectively, and let $a_1 = a \cap A_1$ and $a_2 = a \cap A_2$. The sub-sample of a_2 will be denoted by a_{2m}. Summation over the units in a set s will be denoted by $\sum_s$.

4.3.2.2 The Double Sampling Ratio Estimator in the presence of Non-response

We define the double sampling ratio estimator as follows:

$$\bar{y}_{Rd^*} = \frac{\bar{y}^*}{\bar{x}^*}\bar{x}' = r^*\bar{x}' \tag{4.3.1}$$

Where $\bar{x}^*$ and $\bar{y}^*$ are the Hansen-Hurwitz estimators for $\bar{X}$ and $\bar{Y}$, respectively, and are given by

$$\bar{u}^* = w_1\bar{u}_1 + w_2\bar{u}_{2m}, \qquad u = x, y \tag{4.3.2}$$

According to the general rule, we define $W_j = N_j / N$ and $w_j = n_j / n, j = 1 \text{or } 2$. Sample statistics obtained from a_{2m} are subscripted by "2m", $(e.g., \bar{u}_{2m} = (1/m)\sum_{a_{2m}} u_i)$; those from a_1 are subscripted by "1", $(e.g., \bar{u}_1 = (1/n_1)\sum_{a_1} u_i)$, and those for the first phase sample a' will be superscripted by a prime $(e.g., \bar{x}' = (1/n')\sum_{a'} x_i)$.

A large sample first order approximation to the variance of $\bar{y}_{Rd^*}$, obtained by using the Taylor linearization, is given by

$$V(\bar{y}_{Rd^*}) \cong \left(\frac{1}{n'}-\frac{1}{N}\right)S_y^2+\left(\frac{1}{n}-\frac{1}{n'}\right)S_r^2+\frac{W_2(k-1)}{n}S_{2r}^2 \qquad (4.3.3)$$

Where,

$$S_r^2 = S_y^2 + R^2 S_x^2 - 2RS_{xy},$$

$$S_{2r}^2 = S_{2y}^2 + R^2 S_{2x}^2 - 2RS_{2xy}, \qquad (4.3.4)$$

R is the population ratio of $\bar{Y}$ to $\bar{X}$. S_u^2 and S_{2u}^2 are, respectively, the variance for the stratum of non-respondents of the variable u. S_{xy} and S_{2xy} are the covariance for the whole population and the population of non-respondents respectively.

The variance of $\bar{y}_{Rd^*}$ can be approximately estimated by

$$v(\bar{y}_{Rd^*}) = \left(\frac{1}{n'}-\frac{1}{N}\right)\hat{S}_y^2+\left(\frac{1}{n}-\frac{1}{n'}\right)\hat{S}_r^2+\frac{W_2(k-1)}{n}S_{2r}^2 \qquad (4.3.5)$$

where,

$$\hat{S}_y^2 = \frac{1}{n-1}\left\{\sum_{a_1} y_i^2 + k\sum_{a_{2m}} y_i^2 - n\bar{y}^{*2} + w_{2(k-1)s_{2my}^2}\right\},$$

$$\hat{S}_r^2 = \frac{1}{n-1}\left\{\sum_{a_1}\left(y_i - r^* x_i\right)^2 + k\sum_{a_{2m}}\left(y_i - r^* x_i\right)^2\right\} \text{and}$$

$$\hat{S}_{2r}^2 = \frac{1}{m-1}\sum_{a_{2m}}\left(y_i - r^* x_i\right)^2$$

Note that $\hat{S}_y^2$ is an unbiased estimator of S_y^2.It seems natural to use $\hat{S}_r^2$ to estimate S_r^2 since the expression obtained from $\hat{S}_r^2$ by replacing r^* with R is a consistent estimator of S_r^2. The same argument can be used to justify the use of $\hat{S}_{2r}^2$.

An alternative estimator of $V(\bar{y}_{Rd^*})$ can be obtained by replacing $\hat{S}_r^2$ and $\hat{S}_{2r}^2$ with

$$\tilde{S}_r^2 = \hat{S}_y^2 + r^{*2}S_x'^2 - 2r^*S_{xy}^* \text{ and}$$

$$\hat{S}_{2r}^2 = S_{2my}^2 + r^{*2}S_{2x}^2 - 2r^*S_{2mxy}, \tag{4.3.7}$$

respectively, in (4.3.5), where,

$$s_x'^2 = \frac{1}{n'-1}\sum_{a'}(x_i - \bar{x}')^2, \quad s_{2my}^2 = \frac{1}{m-1}\sum_{a_{2m}}(y_i - \bar{y}_{2m})^2,$$

$$s_{2x}^2 = \frac{1}{n_2-1}\sum_{a_2}(x_i - \bar{x}_2)^2,$$

$$s_{2mxy} = \frac{1}{m-1}\left(\sum_{a_{2m}} x_i y_i - m\bar{x}_{2m}\bar{y}_{2m}\right)$$

and s_{xy}^* is as in (4.3.9). This alternative estimator is likely to have a smaller variance than the estimator in (4.3.5) since the estimators $s_x'^2$ and s_{2x}^2 are based on larger samples and therefore more precise.

4.4.2.3 The Double Sampling Regression Estimator in the presence of Non-response

We define the regression estimator by

$$\bar{y}_{lrt^*} = \bar{y}^* + \hat{\beta}^*(\bar{x}' - \bar{x}^*) \tag{4.3.8}$$

Where $\hat{\beta}^*$ is an estimator of $\beta = S_{xy}/S_x^2$. There could be several choices for $\hat{\beta}^*$, but a natural choice would be given by $\hat{\beta}^* = s_{xy}^*/s_x^{*2}$, where

$$s_{xy}^* = \frac{1}{n-1}\left(\sum_{a_1} x_i y_i + k\sum_{a_{2m}} x_i y_i - n\overline{xy}^*\right)$$

and

$$s_x^{*2} = \frac{1}{n-1}\left(\sum_{a_1} x_i^2 + k\sum_{a_{2m}} x_i^2 - n\overline{xx}^*\right). \tag{4.3.9}$$

It is easy to show that s_{xy}^* and s_x^{*2} are unbiased for S_{xy} and S_x^2 respectively. An approximate variance of $\bar{y}_{lrt^*}$ is given as

$$V(\bar{y}_{lrt^*}) = \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_l^2 + \frac{W_2(k-1)}{n}S_{2l}^2 \tag{4.3.10}$$

Where S_l^2 and S_{2l}^2 are obtained from (4.3.4) by replacing R with β.

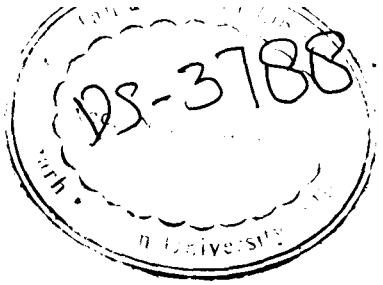
To estimate $V(\bar{y}_{lrt^*})$ we can use the following formula:

$$v(\bar{y}_{lrt^*}) = \left(\frac{1}{n'} - \frac{1}{N}\right)\hat{S}_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)\hat{S}_l^2 + \frac{w_2(k-1)}{n}\hat{S}_{2l}^2 \tag{4.3.11}$$

Where,

$$\hat{S}_l^2 = \frac{1}{n-1}\left\{\sum_{a_1}(y_i - y_i^*)^2 + k\sum_{a_{2m}}(y_i - y_i^*)^2\right\},$$

$$\hat{S}_{2l}^2 = \frac{1}{m-1}\sum_{a_{2m}}(y_i - y_i^*)^2,$$



And

$$y_i^* = \bar{y}_i^* - \hat{\beta}^*(x_i - \bar{x}^*). \tag{4.3.12}$$

Like (2.7), a slightly improved estimator of $V(\bar{y}_{lrt^*})$ can be obtained by using

$$\tilde{S}_l^2 = \hat{S}_y^2 + \hat{\beta}^{*2}s_x'^2 - 2\hat{\beta}^* s_{xy}^*$$

and

$$\tilde{S}_{2l} = s_{2my}^2 + \hat{\beta}^{*2}s_{2x}^2 - 2\hat{\beta}^* s_{2mxy}. \tag{4.3.13}$$

4.3 Choice of Sampling Fractions

We shall now deduce the optimum k, n and n' that minimize the variances of the proposed estimators for a specified cost, or that minimize the cost for a specified variance.

Let's consider a cost function for $\bar{y}_{Rd^*}$ given by

$$C = c'n' + cn + c_1 n_1 + c_2 m \tag{4.4.1}$$

Where the $c's$ are the costs per unit defined as follows:

c': the unit cost associated with the first phase sample, a';

c : the unit cost of the first attempt on y with the second phase sample, a

c_1 : the unit cost for processing the respondent data on y at the first attempt in a_1;

c_2: the unit cost associated with the sub-sample, a_{2m} of a_2.

Since the value of n_1 is not known until the first attempt is made, the expected cost will be used in the minimization.

The expected cost is given by

$$E(C) = C^* = c'n' + \left(c + c_1 w_1 + \frac{c_2 w_2}{k}\right)n. \tag{4.4.2}$$

The optimum value of k, n and n' that minimize the variance of $\bar{y}_{Rd^*}$ for a fixed cost C^* are obtained by using Langrange multiplier. The optimum values thus obtained are:

$$k_o = \sqrt{\frac{c_2\left(S_r^2 - W_2 S_{2r}^2\right)}{S_{2r}^2(c + c_1 W_1)}}$$

$$n_o = \frac{C^*\sqrt{A}}{D\sqrt{G}} \quad \text{and}$$

$$n_o' = \frac{C^*\sqrt{S_y^2 - S_r^2}}{D\sqrt{c'}} \tag{4.4.3}$$

Where

$$A = S_r^2 + W_2(k_o - 1)S_{2r}^2,$$

$$G = c + c_1 W_1 + \frac{c_2 W_2}{k_o} \quad \text{and}$$

$$D = \sqrt{(S_y^2 - S_r^2)c'} + \sqrt{AG}$$

If we let $\gamma = c_2/(c + c_1 W_1)$, $\delta = S_r^2 / S_{2r}^2$ and $\xi = S_y^2 / S_r^2$, then we have

$$k_o = \sqrt{\gamma(\delta - W_2)},$$

$$n_o = \frac{C^* \sqrt{1 + W_2 (k_o - 1)/\delta}}{\sqrt{Gc'(\xi - 1)} + G\sqrt{1 + W_2 (k_o - 1)/\delta}} \quad \text{and}$$

$$n_o' = \frac{C^* \sqrt{\xi - 1}}{c'\sqrt{\xi - 1} + \sqrt{Gc'\{1 + W_2 (k_o - 1)/\delta\}}}. \quad (4.4.4)$$

The optimum values n_o and n_o' are proportional to the expected cost, C^*. To get the optimum values of k, n and n' that, minimize $V(\bar{y}_{lrt^*})$ we simply substitute S_r^2 and S_{2r}^2 in the above expression in (4.4.3) with S_l^2 and S_{2l}^2, respectively.

CHAPTER-V

POST-STRATIFICATION

5.1 Introduction:

In most applications of stratified sampling, the prior knowledge of strata sizes, strata frames and possible variability within stratum are essential requirements. In practical situations, strata sizes are known but lists of stratum units are hard to get. Moreover stratum frames may be incomplete. So one cannot apply stratified sampling for estimating the population parameters. When such type of situation occurs it is used the post-stratification technique is used. Post-stratification means stratification after selection of the sample. The technique consists in selecting a random sample from the entire population and classifying units later according to their representation from different stratums or weights. Usually we only have estimates of the weights based on administrative records, previous census results.

The post-stratification, discussed by Hansen, Hurwitz and Madow (1953). Jager et al. (1985), advocated that with respect to relevant criteria, it may improve upon the estimation strategy subsequently over the sample mean or ratio estimator. Silva and Skinner (1995) used the technique of post-stratification for estimating distribution function with auxiliary information. Shukla and Trivedi (2001, 2006), Shukla et al. (2002) derived methodologies for parameter estimation in sampling under post-stratification.

The mixture of post-stratification and non-response is due to Zhang (1999) who has obtained some results regarding the effect of post-stratification while handling binary survey data subject to non-response. Shukla and Dubey (2001) proposed PSNR (Post Stratified Non Response) sampling scheme for dealing with non-response.

5.2 The Basic Results:

The basic results of post-stratification in case of simple random sampling (SRS) are given below:

If an SRS of size n is drawn from a population of size N, then it is known that the unweighted sample mean

$$\bar{y}_s = \frac{1}{n}\sum_{i\in s} y_i \qquad (5.2.1)$$

is an unbiased estimator of the population mean.

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{N} y_i \qquad (5.2.2)$$

with sampling variance

$$V(\bar{y})_s = \left(\frac{1}{n} - \frac{1}{N}\right)S^2 \qquad (3.2.3)$$

where $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_1 - \bar{Y})^2$

If we have an auxiliary variable X which classifies the sample into k groups with known weights $W_i = \frac{N_i}{N}$, $i = 1,2,...,k$ then the formation of these

groups is called post-stratification. Let the post-stratified data be y_{ij} with $i = 1,2,...,k$ and $j = 1,2,...,n_i$ with post-stratum means given by

$$\bar{y}_i = \frac{1}{n_i} \sum_{j \in s} y_{ij} \tag{3.2.4}$$

Then the post-stratified estimator $\bar{y}_{ps}$ of the population mean

$$\bar{Y} = \sum_{i=1}^{k} \sum_{j=1}^{N_i} y_{ij} / N = \sum_{i=1}^{k} W_i \bar{Y}_i \tag{3.2.5}$$

where

$$\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij} / N_i$$

is given by

$$\bar{y}_{ps} = \sum_{i=1}^{k} W_i \bar{y}_i \tag{3.2.6}$$

Properties of $\bar{y}_{ps}$

We refer to stratified sampling as the conditional distribution, where we condition on the configuration vector

$$n' = (n_1, n_2, ..., n_k) \tag{5.2.7}$$

of the achieved allocation to the strata. If we average over all possible values of n then we obtain the unconditional distribution, which is SRS.

Conditioning on n and using standard results from stratified sampling we have

$$E(\bar{y}_{ps} \mid n) = \sum_{i=1}^{k} W_i E(\bar{y}_i \mid n_i)$$

$$= \sum_{i=1}^{k} W_i \bar{Y}_i = \bar{Y} \qquad (5.2.8)$$

So $\bar{y}_{ps}$ is conditionally unbiased. It then follows that since $\bar{Y}$ is constant it is also unconditionally unbiased. If we now consider the unweighted mean $\bar{y}_s$, we have

$$\bar{y}_s = \sum_{i=1}^{k} w_i \bar{y}_i$$

where $w_i = n_i / n$, and conditionally

$$E(\bar{y}_s \mid n) = \sum_{i=1}^{k} w_i \bar{y}_i \qquad (5.2.9)$$

$$\neq \bar{Y}$$

unless $w_i = W_i$, where $i = 1,2,...,k$. So an estimator $\bar{y}_s$ which is unbiased under SRS is biased under the conditional stratified sampling framework.

In addition to the bias, the distributional framework also embraces the sampling variance. For the conditional distribution, given n, we have

$$V(\bar{y}_{ps} \mid n) = \sum_{i=1}^{k} W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_i^2 \qquad (5.2.10)$$

where

$$S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(y_{ij} - \bar{Y}_i)^2$$

since the conditional distribution given n is equivalent to independent SRS within strata.

The unconditional variance is

$$V(\bar{y}_{ps}) = \sum_{i=1}^{k} W_i^2 \left(E\left(\frac{1}{n_i}\right) - \frac{1}{N_i} \right) S_i^2$$

$$\approx \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i S_i^2 + \frac{1}{n^2}\sum_{i=1}^{k}(1 - W_i) S_i^2 \qquad (5.2.11)$$

The first term is the variance of a proportionally allocated stratified sample while the second term is often small. Thus, on average, the gains from post-stratification are asserted to be about the same as those of a proportionally allocated stratified sample. Since in social surveys these gains tend to be rather small, the value of post-stratification has often been questioned. However, major gains can be obtained in institutional and business surveys, and it is here that post-stratification is widely employed.

Holt & Smith(1979) show that, on average, these statements can be very misleading for inference. By considering the coverage properties of both the conditional and unconditional confidence intervals, they show that the conditional intervals obtained using $\bar{y}_{ps}$ and $V(\bar{y}_{ps} \mid n)$ in equation (5.2.10) are reasonable for any configuration n. On the other hand, the unconditional intervals based on $\bar{y}_{ps}$ and $V(\bar{y}_{ps})$ in equation (5.2.3) do not contain $\bar{Y}$ with the correct frequency for most configurations. On average, they give

the right coverage but conditionally they are inadequate. Since once the sample is drawn the configuration n is known, it seems appropriate to make inferences which give adequate coverage conditional on n.

5.3 Efficient Estimation in Post Stratified Non Response Sampling Scheme Using Auxiliary Information

The non-response in the mail surveys is a kind of incompleteness that occurs when respondents do not reply through their mails and the sample remains incomplete. In the set-up of stratified sampling, when stratum sizes are unknown, the post stratification is a useful strategy (see, Cochran (1999), Mukhopadhyay (1999) etc.). The non-response in surveys could be handled by various techniques and methodologies [see Lessler and Kalsbeek (1992), Grover and Couper (1998), Khare (1987) and Khot (1994)] and one of them is the imputation of available data as described by Hinde and Chambers (1991).

Let a population of size N be divided into k strata. The size of the $i-th$ strata be $N_i, (i=1,2,...,k)$, such that $\sum_{i=1}^{k} W_i = \sum_{i=1}^{k} \frac{N_i}{N} = 1$.

In the following we give symbols, which are commonly used-

N_i' : Number of respondents (RS) in $i-th$ strata.

N_i'' : Number of non respondents (NRS) in $i-th$ strata.

$\bar{Y}_i'$: Population mean of $i-th$ strata for response group (R) for Y.

$\bar{Y}_i''$: Population mean of $i-th$ strata for non-response group (NR) for Y.

$\bar{Y}$: Population mean for Y, i.e. $\bar{Y} = \sum_{i=1}^{k} W_i \bar{Y}_i$.

S_{1iY}^2 : Population mean square in $i-$th strata for response group (R) for Y.

S_{2iY}^2 : Population mean square in $i-$th strata for non-response group (NR) for Y.

C_{1iY}: Coefficient of variation in $i-$th strata for response group (R) for Y.

C_{2iY}: Coefficient of variation in $i-$th strata for non-response group (NR) for Y.

We assume X an auxiliary variable, correlated with Y.

$\bar{X}_i'$: Population mean in $i-th$ strata for response group (R) of X.

$\bar{X}_i''$: Population mean in $i-th$ strata for non-response group (NR) of X.

S_{1iX}^2 : Populations mean squre in $i-th$ strata for response group (R) for X.

S_{2iX}^2 : Population mean square in $i-th$ strata for non-response group (NR) for X.

C_{1iX} : Coefficient of variation in $i-th$ strata for response group for X.

C_{2iX} : Coefficient of variation in $i-th$ strata for non-response group for X.

ρ_i' : Correlation coefficient for response group (between X_i' and Y_i').

ρ_i'' : Correlation coefficient for non-response group (between X_i'' and Y_i'').

R_i' : Ratio in $i-th$ strata for response group, i.e. $R_i' = \frac{\bar{Y}_i'}{\bar{X}_i'}$.

R_i'' : Ratio in $i-th$ strata for non-response group, i.e. $R_i'' = \frac{\bar{Y}_i''}{\bar{X}_i''}$.

We note that $N_i = N_i' + N_i''$ and $n_i = n_i' + n_i''$.

5.4 Post Stratified Non-Response (PSNR) Sampling Scheme With Auxiliary Variable:

PSNR sampling scheme is described into following steps:

Step I: Select a sample of size n by SRSWOR from the population N and post-stratified into k strata, such that n_i units represent to N_i $\left(\sum_{i=1}^{k} n_i = n\right)$.

An auxiliary source of information (other than X) may be used for this purpose.

Step II: Mail questionnaires to all the random n_i units for response over the variable Y under study and wait until a deadline. If possible complete response occurs, $\bar{y}_i$ is sample mean from $i-th$ strata and $\bar{y} = (n)^{-1}\sum_{i=1}^{k} n_i\, \bar{y}_i$.

Step III: Assume that non-response observed when the deadline of returning questionnaire is over, and there are n_i' respondent, n_i'' non-respondents in the $i-th$ strata$(n_i' + n_i'' = n_i)$. The $\bar{y}_i'$ is mean of responding n_i' units for the

Y and $\bar{x}_i'$ is mean of corresponding n_i' units over auxiliary variable X. Moreover, $R_i = \frac{Y_i}{X_i}$ is the true ratio in the stratum i and it is assumed that respondents for Y must have responded for X also among n_i'.

Step IV: From non-responding n_i'', select sub-samples of size n_i''' by SRSWOR, maintaining a prefixed fraction $f_i = \left(\frac{n_i'}{n_i'''}\right)$ over all the k strata.

Step V: Conduct a personal interview for n_i''' units and assume all these responded well during that period over Y and X both. The $\bar{y}_i'''$ is the mean based on n_i''' and $\bar{x}_i'''$ is the corresponding mean of X.

5.5 The Proposed Estimator:

The proposed estimation strategy is

$$\bar{y}_{rPSNR} = \sum_{i=1}^{k} W_i \left[\left(\frac{n_i'}{n_i}\right)(\bar{y}_{ir}') + \left(\frac{n_i''}{n_i}\right)(\bar{y}_{ir}''')\right],$$

where *rPSNR* stands for "ratio post-stratified non-response", $\bar{y}_{ir}'$ is the ratio estimate for mean of responding units of $i-$th strata in stratified random sampling [*i.e.* $\bar{y}_{ir}' = \frac{\bar{y}_i'}{\bar{x}_i'} \bar{X}_i'$] and $\bar{y}_{ir}'''$ is the ratio estimate for mean based on n_i''' units in stratified random sampling [*i.e.* $\bar{y}_{ir}''' = \frac{\bar{y}_i'''}{\bar{x}_i'''} \bar{X}_i''$], where $\bar{X}_i'$ and $\bar{X}_i''$ is assumed to be known.

Theorem 5.5.1: $\bar{y}_{rPSNR}$ is biased for $\bar{Y}$ and the approximate amount of bias is

$$Bias(\bar{y}_{rPSNR}) = \sum_i W_i \left(\frac{N_i'}{N_i}\right) \bar{Y}_i' \left(\frac{N-n}{Nn}\right) (C_{1iX}^2 - \rho_i' C_{1iX} C_{1iY})$$

$$+ \sum_i W_i \left\{\frac{1}{nW_i} + \frac{(N-n)(1-W_i)}{n^2(N-1)W_i^2}\right\} \bar{Y}_i'' (f_i - 1)(C_{2iX}^2 - \rho_i'' C_{2iX} C_{2iY})$$

Remark 5.5.1:

We use some standard results in the proof of above expression

i) $$E\left[\left(\frac{n_i'}{n_i}\bar{y}_{ir}'\right)|n_i\right] = \frac{N_i'}{N_i}\bar{Y}_i'\left[1 + \frac{N_i' - n_i'}{N_i' n'}(C_{1iX}^2 - \rho_i' C_{1iX} C_{1iY})\right]$$

ii) $$E\left[\left(\frac{n_i''}{n_i}\bar{y}_i''\right)|n_i\right] = \frac{N_i''}{N_i}\bar{Y}_i''$$

iii) $$E\left(\frac{1}{n_i}\right) = \left[\frac{1}{nW_i} + \frac{(N-n)(1-W_i)}{n^2(N-1)W_i^2}\right]$$

Theorem5.5.2: The approximate expression of variance of $\bar{y}_{rPSNR}$ is

$$V(\bar{y}_{rPSNR}) = \sum_{i=1}^{k} W_i^2 \left[\left(E\left(\frac{1}{n_i}\right)\right)\left(\frac{N_i''}{N_i}\right)(f_i - 1)A_i'\right] + \sum_{i=1}^{k} W_i^2$$

$$\left\{E\left(\frac{1}{n_i}\right)W_i(1-W_i)\right\} \left[\left(\frac{N-n}{Nn}\right)B_i' + \left\{1 + \left(\frac{N-n}{Nn}\right)C_i'\right\}^2 \left\{\left(\frac{1}{n} - \frac{1}{N}\right)S_{2iy}^2\right\}\right],$$

where

$$A_i' = [S_{2iy}^2 + R_i''^2 S_{2ix}^2 - 2\rho_i'' R_i'' S_{2iy} S_{2ix}]$$

$$B_i' = [S_{1iy}^2 + R_i'^2 S_{1ix}^2 - 2\rho_i' R_i' S_{1iy} S_{1ix}]$$

$$C_i' = (C_{2ix}^2 - \rho_i'' C_{2iy} C_{2ix}).$$

5.6 Cost Analysis:

Cost analysis is incorporated to estimate the optimal sample size, consider $i-th$ stratum-based approach by assuming cost C_{0i}, C_{1i}, C_{2i} varying over all k strata.

C_{0i} : cost of n_i units of the $i-th$ stratum.

C_{1i} : cost of collecting, editing and processing per n_i' units in the $i-th$ strata of response class.

C_{2i} : cost of personal interview and processing per n_i''' units in the $i-$th strata for non-response class.

The $i-$th stratum has total cost

$$C_i = [C_{0i} n_i + C_{1i} n_i' + C_{2i} n_i'''] \qquad (5.6.1)$$

Total cost over k strata

$$T_c = \sum_{i=1}^{k} C_i = \sum_{i=1}^{k} [C_0 n_i + C_1 n_i' + C_2 n_i''']$$

$$E(T_c) = \left(\frac{n}{N}\right) \sum_{i=1}^{k} \left\{ C_0 N_i + C_1 N_i' + C_2 \frac{N_i''}{f_i} \right\}$$

To get optimum f_i, λ_i and n, define a function L_i, with lagrange mulitiplier λ_i and pre-fixed level of variance V_{0i}

L_i = [expected cost of $i-th$ strata] + $\lambda_i[V(\bar{y}_{rPSNR})_i - V_{0i}]$

$$L_i = \left(\frac{n}{N}\right)\left[C_{0i} N_i + C_{1i} N_i' + C_{2i}\left(\frac{N_i''}{f_i}\right)\right] + \lambda_i \left[\left\{ \left(E\left(\frac{1}{n_i}\right)\right)\right\}\left(\frac{N_i''}{N_i}\right)(f_i - 1)A_i' \right]$$

$$+ \lambda_i \left\{ E\left(\frac{1}{n_i}\right) W_i(1-W_i) \right\} \left[\left(\frac{N-n}{Nn}\right) B_i' + \left\{ 1 + \left(\frac{N-n}{Nn}\right) C_i' \right\}^2 \right.$$

$$\left. \left\{ \left(\frac{1}{n} - \frac{1}{N}\right) S_{2iy}^2 \right\} \right] - \lambda_i V_{0i} \qquad (5.6.2)$$

On differentiating (5.6.2) with respect to f_i, λ_i and n, we get three equations.

Differentiating with respect to f_i, we get

$$-\left(\frac{n}{N}\right) \frac{C_{2i}}{f_i^2} N_i'' + \lambda_i \left[E\left(\frac{1}{n_i}\right)\left(\frac{N_i''}{N_i}\right) A_i' \right] = 0 \qquad (5.6.3)$$

To simplify the expression further, we shall use approximation

$$\frac{N_i'}{N_i} = \frac{N_i''}{N_i} = \frac{N_i}{N} \cong W_i$$

$$\text{or } \lambda_i = (n\, C_{2i}\, W_i) \left\{ f_i^2 E\left(\frac{1}{n_i}\right) A_i' \right\}^{-1} \tag{5.6.4}$$

Secondly, differentiating (5.6.2) with respect to λ_i, we get

$$\left[\left\{E\left(\frac{1}{n_i}\right)\right\}\left(\frac{N_i''}{N_i}\right)(f_i - 1)A_i'\right] + \left\{E\left(\frac{1}{n_i}\right)W_i(1-W_i)\right\}$$

$$\left[\left(\frac{N-n}{Nn}\right)B_i' + \left\{1+\left(\frac{N-n}{Nn}\right)C_i'\right\}^2 \left\{\left(\frac{1}{n}-\frac{1}{N}\right)S_{2iy}^2\right\}\right] - V_{0i} = 0 \tag{5.6.5}$$

Thirdly, differentiating (5.6.2) with respect to n, and neglecting the terms of order n^{-3}, we get

$$[C_{0i}W_i + C_{1i}W_i] + C_{2i}\frac{W_i}{f_i} - \left(\frac{\lambda_i}{n}\right)\left[\left\{E\left(\frac{1}{n_i}\right) + \frac{N(1-W_i)}{n^2(N-1)W_i^2}\right\}W_i(f_i-1)A_i'\right]$$

$$-\left(\frac{\lambda_i}{n^2}\right)\left[\left\{2E\left(\frac{1}{n_i}\right) + \frac{N(1-W_i)}{n^2(N-1)W_i^2}\right\}W_i(1-W_i)\left(S_{2iy}^2 + B_i'^2\right)\right] = 0 \tag{5.6.6}$$

Putting the value of λ_i from (5.6.4) in (5.6.6), we get

$$l_{1i}\, f_i^2 - l_{2i}\, f_i + l_{3i} = 0 \tag{5.6.7}$$

$$\text{or } f_i = \left(\frac{1}{2l_{1i}}\right)\left(l_{2i} \pm \sqrt{l_{2i}^2 - 4l_{1i}l_{3i}}\right), \tag{5.6.8}$$

where

$$l_{1i} = (C_{0i} W_i + C_{1i} W_i), \qquad l_{2i} = \left(C_{2i} W_i - \frac{C_{2i} W_i^2 \alpha_i}{E(1/n_i)} \right)$$

$$l_{3i} = \left(\frac{C_{2i} \alpha_i W_i^2}{E(1/n_i)} - \frac{C_{2i} W_i G_i}{E(1/n_i) A_i'} \right), \quad \alpha_i = E\left(\frac{1}{n_i} \right) + \frac{N(1-W_i)}{n^2 (N-1) W_i^2},$$

$$G_i = \left\{ 2E\left(\frac{1}{n_i} \right) + \frac{N(1-W_i)}{n^2 (N-1) W_i^2} \right\} W_i (1-W_i) \left(\frac{S_{2iy}^2 + B_i'}{n} \right).$$

The positive value of f_i will occur if the condition $l_{2i}^2 > (l_{1i} l_{3i})$ holds. On substituting the standard values of $E\left(\frac{1}{n_i} \right)$ in (5.6.5), n optimum can be obtained from the following equation:

$$n^2 V_{0i} - n E_i - F_i = 0, \tag{5.6.9}$$

where

$$E_i = \frac{P_i}{W_i} - \frac{(1-W_i) P_i}{(N-1) W_i^2}, \quad F_i = \frac{(1-W_i) P_i}{W_i^2} + \frac{Q_i}{W_i},$$

$$Q_i = W_i (1-W_i) \left(B_i' + S_{2iy}^2 \right), \quad P_i = W_i (f_i - 1) A_i'.$$

This equation gives the optimum $(n_{opt})_i$ using f_i from (5.6.9) along with a suitable choice of V_{0i}.

REFERENCES

Cochran, W. G. (1961): Comparison of methods for determining stratum boundaries. *Bull. Int. Stat. Inst.*, **38, 2,** 345-358.

Cochran, W. G. (1963): *Sampling Techniques*, John Wiley, New York.

Cochran, W. G. (1977): *Sampling Techniques*, John Wiley, New York.

Delenius, T. (1950): Problem of optimum stratification, *Skandinavisk Aktuarietidskrift*, **33**, 203-211.

Delenius, T., and Gurney, M. (1951): The problem of optimum stratification II. *Skandinavisk Aktuarietidskrift*, 34, 133-148.

Hansen, M. H., and Hurwitz, W. N. (1946): The problem of response in sample survey, *J. Amer. Statist. Assoc*, **41**, 517-529.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953): *Sample survey method and theory*, 2, John Wiley, New York.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953): *Sample Survey Methods and Theory*. John Wiley, New York.

Hayashi, C. Maruyama, F. and Isida, M. D. (1951): On some criteria for statistification, *Ann. Ins. Statist. Math.*, **2,** 77-86.

Hinde, R.L. and Chambers, R.L. (1991): Non response imputation with multiple source of non-response. *J. Official Statist.*, 7, 169-179.

Jessen, R. J. (1942): Statistical investigations of a sample survey for obtaining farm fact, Iowa Agricultural Experimental Station, *Research Bulletin* **304.**

Kalton, G., and Kasprzyk, D. (1886): The treatment of missing survey data,*Survey Methodology*, **12**, 1-16.

Khae, B.B. (1987): Allocation in stratified sampling in presence of non-response. *Metron*, **XLV**, 213-221.

Khot, P.S.(1994). A note on handling non-response in sample Surveys. *J. Amer. Statist. Assoc.*, **89**, 693-696.

Lessler, J.T. and Kalsbeek, W.D. (1992): *Non-response Error in Surveys.* John Wiley, New York.

Mahajan, P. K. Gupta, J. P. and Singh, R. (1994): Determining of optimum strata boundaries for scrambles randomized response. *Statistica, Anno.*, **3**, 375-381.

Mahalanobis, P. C. (1952): Some aspects of the design of sample surveys. *Sankhyā* , **12**, 1-7.

Mukhopadhyay, P. (1999): *Theory and Methods of Survey Sampling.* Prentice-Hall of India, New Delhi.

Patterson, H. D. (1950): Sampling on successive occasions with partial replacement of units, *J. Roy. Statist. Soc.*, **B12**, 241-255.

Sen, A. R. (1971): Increased precision in canadian waterfowl harvest survey through successive sampling, *J. Wild Manag.*, **35**, 664-668.

Seti, V. K. (1963): A note on optimum stratification for estimating the population means. *Aust. J. statist.*, **5**, 20-33.

Shukla, D., Bankey, A. and Trivedi, M. (2002): Estimation in post stratification using prior information and grouping strategy. *J.Indian Soc. Agricultural Statist.*, **55**, 158-173.

Shukla, D. and Dubey, J. (2001): Estimation in mail surveys under *PSNR* sampling scheme. *J. Indian Soc. Agricultural Statist.*, **54**, 288-302.

Shukla, D. and Dubey, J. (2004): On estimation under post-stratified two phase non-response sampling scheme in mail surveys. *Int. J. Manag. Syst.*, **20,** 269-278.

Shukla, D. and Dubey, J. (2006): On earmarked strata in post-Stratification". *Statist. Transition*, **7**, 1067-1085.

Shukla, D. and Trivedi, M. (2001): Mean estimation in deeply stratified population under post-stratification". *J. Indian Soc. Agricultural Statist.*, **54**, 221-235.

Shukla, D. and Trivedi, M. (2006): Examining stability of variability ratio with application in post-stratification. *Int. J. Manag. Syst.*, **22**, 59-70.

Silva, P.L.D.N. and Skinner, C.J.(1995): Estimation of distribution functions with auxiliary information using post stratification". *J.Official Statist.*, **3**, 277-294.

Singh, R. (1977): A note on optimum stratification for equal allocation with ratio and regression methods of estimation. *Aust. J. Stat.* **19**, 96-104.

Singh, R. and Sukhatme, B. V. (1969): Optimum stratification, *Ann. Ins. Statist. Math,* **21**, 515-528.

Singh, R. and Sukhatme, B. V. (1973): Optimum stratification with ratio and regression methods of estimation, *Ann. Ins. Statist. Math.*, **25**, 627-633.

Smith, T.M.F. (1991): Post-stratification. *The Statisticians*, **40**, 323-330.

Sukhatme, P.V., Sukhatme, B.V., Shukatme, S. and Ashok, C. (1984): *Sampling Theory of Surveys with Applications*. Iowa State University Press, U.S.A.

Wywial, J. (2001): Stratification of the population after sample Selection. *Statistics in Transition*, **5**, 327-348.

Yates, R. (1960): *Sampling Methods for Censuses and Survey.* Griffin, London.

Zhang, L.C. (1999): A note on post-stratification when analyzing binary survey data subject to non-response. *J. Off. Statist.*, **15**, 329-334.